1. **The Climate Survey Design Project: One Solution**

An optimization problem was presented in Chapter 2 for a single-stage stratified sample design. In the following sections, we present a solution to the multipurpose design question borrowing from material presented in Chapters 3-6. A series of solutions was generated for the sample allocation to test the sensitivity of the assumptions. Additionally, different software may produce different yet comparable results. Ultimately, a single solution must be chosen from this set for implementation as discussed below.

* 1. **Overview of the Project**

The Senior Council within the Verkeer NetUltraValid (VNUV) Corporation has tasked your design team with developing an optimal allocation for their annual employee climate survey—the VNUV Climate Survey, Cycle 5. The survey sample members will be randomly selected through a single-stage stratified design as employed in previous VNUV climate surveys. The analysis variables of interest for the survey include:

1. (Q5) Overall, I am satisfied with VNUV as an employer at the present time.
2. (Q12) There is a clear link between my job performance and my pay at VNUV.
3. (Q15) Overall, I think I am paid fairly compared with people in other organizations who hold jobs similar to mine.
4. The number of training classes attended by the employees in the past 12 months.

The design team met over a three-week period to develop the sample design. During this period, they

1. finalized the assumptions used for the optimization;
2. formulated the optimization problem;
3. constructed and implemented computer programs to obtain multiple solution;
4. developed a presentation to highlight the results to the Senior Council (not shown); and
5. summarized the work in a final report (not shown).
   1. **Formulate the Optimization Problem**

The first task for mathematical modeling as discussed in Chapter 5 is to translate the client’s needs and constraints for a survey into a set of equations that can be solved. This is similar to the task of translating word problems into equations in our first algebra class. Following the components discussed in Section 5.1, we extract the necessary information from Chapter 2 to construct the multicriteria optimization problem.

* + 1. **Objective Function**

The objective function is the equation that is minimized or maximized to develop a solution. Skimming back through Chapter 2, you will not locate an explicit definition for this function. Welcome to one of the many areas where creativity plays a role in the lives of survey statisticians. Through experience you may develop a preference for a particular type of objective function. Otherwise, the use of more than one objection function (and set of assumptions) can suggest the robustness of your final chosen solution.

Based on previous experience, the objective chosen by the design team was similar to the equation used for Example 5.2. Namely, the allocation should be constructed to minimize the sum of the relvariance of the estimated total () for the four analysis variables (Section 2.1; repeated in Section 7.1 for convenience). In other words, the explicit formula for the first candidate objective function is

,

where  is the importance weights for variable *j* (),  is the corresponding relvariance such that , and  is the unit variance calculated within design stratum *h* (). The design team had several discussions about the importance weights used in the objective function. After conferring with the Senior Council, the decision was reached that all of the analysis variables were of equal importance. Consequently,  for all four variables so that expression is rewritten as

.

Several objective functions could have been tested. However, because of the time commitments for the design time (a common constraint for researchers), the objective function discussed in Chapter 5 was borrowed for this project.

* + 1. **Decision Variables**

The decision variables correspond to the solutions produced from the optimization problem, i.e., sample size and associated allocation to strata. For the VNUV Climate Survey, the allocation solution for the 18 design strata – business unit (3) by salary grade (3) by employment tenure (2) – as shown in Table 2.1. Note that the solution is derived to meet certain analytic objectives specified for the survey. Once the solution has been obtained, the values must be inflated to address sample loss associated with study ineligibility and nonresponse (Chapter 6).

* + 1. **Optimization Parameters**

Three sets of parameters were defined for the optimization problem. First, HR provided counts of eligible employees by the sampling strata. These frame counts are shown in Table 2.1 of Chapter 2. Second, the design team incorporated the performance rates calculated from the Cycle 4 study (Table 2.3) to ensure that that the analytic objectives could be met with the total number of respondents as well as their distribution across the sampling strata. The last set of parameters includes the population estimates, means/proportions and standard errors, shown in Table 2.4. Prior to implementation, the design team constructed population standard deviations from the estimated standard errors using the expression (3.39):



Note that  for the Cycle 5 calculations because the sample for the previous climate survey was selected by an *srs* design. We visit the design effect again for the Cycle 5 design in section 7.3.

* + 1. **Specified Survey Constraints**

Questions were posed to the VNUV Senior Council to finalize the optimization constraints on the sample size and on the precision for a set of estimates (Section 2.2). The first constraint was dictated by the survey budget – there are sufficient funds for the Cycle 5 climate survey to process responses from 600 sample members. In addition to constraining the sum of the respondent sizes generated from the allocation, the design team also required that the number in each stratum exceed a specified minimum value in order to calculate a variance component. Because the actual number selected for the study was calculated as the respondent size inflated for sample loss (e.g., nonresponse) determined from the Cycle 4 survey, the inflated size was constrained to be less than the frame count within the stratum. In summary, the following set of equations was used to constrain the sample allocation:





where *nh* is the number of respondents within stratum *h*  derived from the optimum allocation, *Nh* is the total number of employees in stratum *h* calculated from the updated employee list provided by HR (see Table 2.1), and *rh* is the sample-loss inflation rate from Cycle 4 calculated as the eligibility rate (= 1 – ineligibility rate) multiplied by the response rate (see Table 2.3).

A second set of constraints was placed on the coefficient of variation (CV) for four estimates (Q5, Q12, Q15, and the average number of training classes) within domains defined by business unit, salary grade within business unit, and categorized tenure within business unit (Table 2.6).

A third set of constraints were imposed by the design team prior to finalizing optimization. These constraints were derived from a power analysis discussed in the next section.

* 1. **One Solution**
     1. **Power Analyses**

Having specified the known constraints for the optimization task, the design team next conducted a power analysis to establish a minimum sample size for the business unit domains to meet the desired detectable differences:

* a five percentage point difference (or larger) for the employee climate estimates, and
* a difference of two to three training classes for the average on-the-job education estimates.

The design team, however, eventually determined from the power analysis that the desired difference levels were not attainable given the study budget, i.e., the funds used to edit and analyze data from 600 respondents.

The multivariable power analysis focused on four estimates. Beginning with the proportion of staff who (strongly) agrees with the three climate questions restated in section 7.1, Table 2.5 in Chapter 2 showed that the fair compensation question (Q15) consistently had the lowest rate of agreement across the business units. The design team noted that the Q15 estimate had the strongest influence on the power calculations because it has the largest standard deviation. Thus Q5 and Q12 were set aside and not used in the minimum sample size analysis. The influence of Q15 in comparison of the average number of training classes was less clear so two separate power calculations were made and subsequently combined.

The R function power.prop.test produced the results shown in Table 7.1. For example, the R code used to calculate the sample size for the SR business unit with a detectable difference of 0.05 (or delta=5 percentage points) is

power.prop.test(p1=0.69, p2=0.74, sig.level = 0.05,

power = 0.8, alternative = "two.sided").

Similar code was used to calculate the minimum analytic sample size for the CR and FO business units. Note that each value in Table 7.1 from the power analysis with delta=5 percentage points violates the constraint of 600 respondents. The team reran the analysis using several detectable differences; power results for 10, 13, and 15 percent are included in the table for comparison. The values for 0.13 looked most promising because the total

**Table 7.1** Minimum sample size by business unit and detectable difference produced by the R function power.prop.test for the fair salary question (Q15).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Business | Q15 Cycle-4 | Detectable Difference (%) | | | |
| Unit | Estimate | 5 | 10 | 13 | 15 |
| SR | 0.69 | 1,278.3 | 300.8 | 170.6 | 124.2 |
| CR | 0.83 | 777.3 | 164.6 | 86.1 | 58.8 |
| FO | 0.60 | 1,470.5 | 355.9 | 205.7 | 151.9 |
| Overall |  | 3,526.1 | 821.4 | 462.5 | 334.9 |

a 80 percent power and 0.05 level of significance for a two-sided test.

**Table 7.2** Minimum sample size by business unit and detectable difference produced by the R function power.t.test for the number of trainings question.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Business | Cycle-4 Estimates | | | Detectable Difference (Trainings) | | | |
| Unit | Mean | *se* | *std* | 1.0 | 1.5 | 2.0 | 2.5 |
| SR | 18.10 | 0.98 | 12.02 | 1,036.5 | 461.8 | 260.6 | 167.5 |
| CR | 12.60 | 0.90 | 8.21 | 491.3 | 219.4 | 124.3 | 80.2 |
| FO | 8.94 | 0.60 | 7.74 | 431.5 | 192.9 | 109.3 | 70.7 |
| Overall |  |  |  | 1,959.3 | 874.0 | 494.2 | 318.4 |

a 80 percent power and 0.05 level of significance for a two-sided test.

**Table 7.3** Minimum sample size by business unit for design optimization of the VNUV Climate Survey, Cycle 5.

|  |  |  |
| --- | --- | --- |
| Business  Unit | Minimum #  Respondents | DEFF Adjusted #  Respondentsa |
| SR | 170.6 | 179.2 |
| CR | 86.1 | 90.4 |
| FO | 205.7 | 216.0 |
| Overall | 462.5 | 485.6 |

a Design effect of 1.05 used to account for variation introduced through weighting.

sample size was well below the maximum value and would hopefully allow the optimization algorithm some flexibility in allocating sample across the strata. Next, the team turned to a similar calculation for the average number of training classes.

The team accessed R again to calculate the minimum sample size per business unit for the average number of training classes with the power.t.test function. Table 7.2 contains the results from the second power analysis for a range of days included as the desired detectable difference. The standard deviations (*std*) were calculated with expression with . Detectable differences between 2 and 3 classes were classified by the Senior Council as meaningful. Differences less than these numbers were examined to evaluate the sample size requirements for higher levels of precision.

Having examined the results, the team decided to take the “best of both worlds.” The maximum sample size required by business unit for a 13 percentage point difference in the climate estimates and a 2.5 difference in the average number of training days combines the information, resulting in the values given in the “minimum # respondents” column of Table 7.3.

Because the Cycle 5 post-data collection analysis will include the use of weights, unlike Cycle 4, a senior statistician was consulted on an appropriate design effect. A conservative *deff* of 1.05 was used to inflate the analytic sample size to account for factors such as differential weights introduced from a nonresponse adjustment. These inflated values located in the last column of the table were used in the optimization routines discussed next.

* + 1. **Optimization Results**

The sample allocation was optimized using both Excel Solver and SAS proc optmodel for comparison. The output files from the optimizations are located on the book’s website as discussed below.

Solver. The file containing the Solver output is named *Project 1.Solver.xlsx*. The workbook contains 14 worksheets, some corresponding to tabular information provided in Chapter 2:

* 1 Frame Counts (Table 2.1)
* 2 Recode (Table 2.2)
* 3 Study Rates (Table 2.3)
* 4\_5 Estimates (Tables 2.4 and 2.5)
* 6 CVs (Table 2.6)

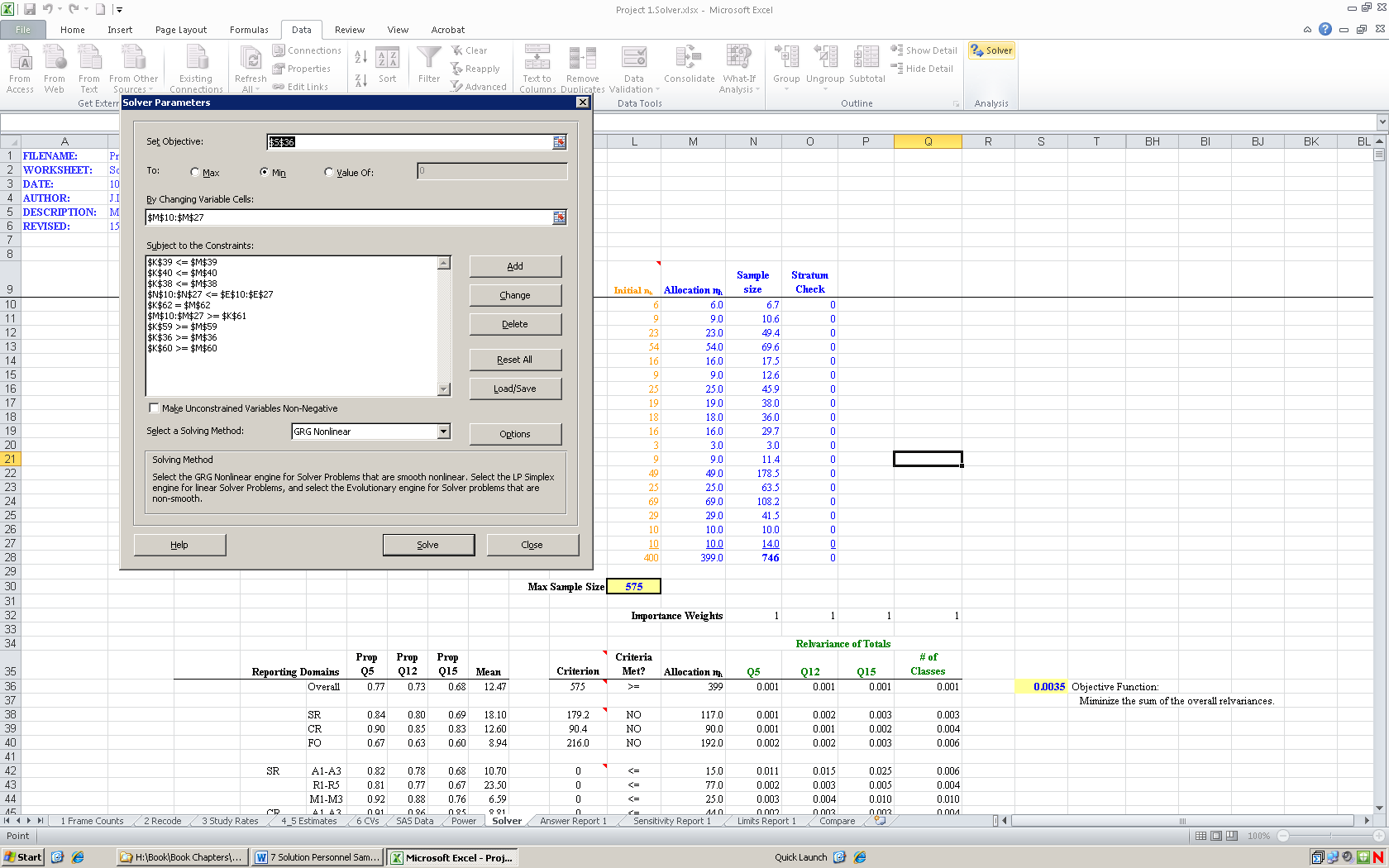
and some containing input or output from the optimizations:

* SAS Data (input data for SAS proc optmodel)
* Power (summary tables from R power functions)
* Solver (Excel Solver optimization)
* Answer Report 1, Sensitivity Report 1, Limits Report 1 (output from Solver)
* Answer Report 2 (output from Solver with multistart option)
* Compare (comparison between Solver and proc optmodel solutions)
* Sensitivity (sensitivity of Solver solution to changes in the assumed response rates)

Details from the Solver optimization are summarized below:

1. The default settings for Solver were used including GRG Nonlinear solving method, 0.0001 precision constraint, and 1,000 iterations. The optimization was computed both with and without the “multistart” option resulting in no difference.
2. As shown in Figure 7.1, the objective function, equation , is tabulated in cell S36 within the Solver worksheet (or ‘Solver’!$S$36 using Excel notation). The goal of the optimization is to minimize the sum of the four relvariances, one for each estimate. The change cells, or the respondent sample size per strata, are located in column M, rows 10 through 27. The series of constraints have been loaded into the third input box included the maximum sample size ($K$36 >= $M$36), the minimum sample size per business unit determined from the power calculations ($K$38 <= $M$38 through $K$40 <= $M$40), and an additional check to ensure that the allocation inflated for sample loss would not exceed the frame counts per strata ($N$10:$N$27 <= $E$10:$E$27).

**Figure 7.1** Excel Solver optimization parameter input box



1. A proportional allocation was used for the starting values— see 'Solver'!L9.
2. The original optimization was implemented using a maximum respondent sample size of 600. Because the constraints were easily met, the team evaluated a reduced respondent sample size in an attempt to save project time and funds. The final recommended sample size was 575 respondents.[[1]](#footnote-1)

SAS proc optmodel. The SAS programs, logs, and lst (output) files are identified by the label “Project 1 OptModel n=\*” where the asterisk (\*) indicates the maximum respondent sample size set for the optimization routine. The SAS programmer included the sample size constraints overall and by business unit along with the CV constraints as macro variables at the beginning of the program. Each section of the program either inputs tables specified in Chapter 2 or calculates components for the optimization. The SQP (default) procedure was used in the optimization as shown in the log files.

The design team initially produced an optimization for a 600 respondents to mirror the work completed with Solver (see the files “Project 1 OptModel n=600”). Two additional SAS programs were created corresponding to n=575 and n=550 respondents. As shown in Project 1 OptModel n=500.log, a feasible solution with a maximum of 550 respondents was not found.

Comparison of Solver and SAS proc optmodel. A comparison of the allocation results from Excel Solver and SAS proc optmodel is shown in Table 7.4. The *solution* to both algorithms satisfied the revised respondent sample size of 575. However, after inflating the allocation for sample loss (*adjusted solution*) and randomly rounding[[2]](#footnote-2) the adjusted values (*random round*), the Solver solution required the selection of 36 (=1,025 – 1,061) fewer sample cases. Additionally, the Solver objective function was 4 percentage points lower (4.82 vs. 4.87), indicating a more precise solution. Therefore, the design team chose the Solver solution as the sample allocation included in the report.

* 1. **Additional Sensitivity Analysis**

The design team completed one last analysis prior to finalizing the report to the VNUV Senior Council to address the concerns about the estimated response rates (see the response to question #8 in section 2.2). Without detailed information on the likely differential rates by the stratifying characteristics, the team evaluated the impact of an overall reduction in the response rates to identify subgroup estimates that would be most affected. In summary, that group is the Survey Research (SR) division. The following three points are the take-away messages:

**Table 7.4** Comparison of optimization results from Excel Solver and SAS proc optmodel for the VNUV Climate Survey, Cycle 5.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Strata** | **Business Unit** | **Salary Grade** | **Tenure** |  | **Solver** | | |  | **OptModel (SQP)** | | |
|  | **Solution**1 | **Adjusted Solution**2 | **Random Round**3 |  | **Solution** | **Adjusted Solution**1 | **Random Round** |
| 1 | SR | A1-A3 | <5 Yrs |  | 12.7 | 14.3 | 15 |  | 10.9 | 12.3 | 12 |
| 2 |  |  | 5+ Yrs |  | 20.1 | 23.8 | 23 |  | 17.8 | 21.0 | 21 |
| 3 |  | R1-R5 | <5 Yrs |  | 34.1 | 73.2 | 74 |  | 33.5 | 71.9 | 71 |
| 4 |  |  | 5+ Yrs |  | 65.1 | 83.8 | 83 |  | 81.4 | 104.8 | 104 |
| 5 |  | M1-M3 | <5 Yrs |  | 27.1 | 29.7 | 30 |  | 23.1 | 25.3 | 26 |
| 6 |  |  | 5+ Yrs |  | 20.1 | 28.1 | 28 |  | 12.5 | 17.5 | 17 |
| 7 | CR | A1-A3 | <5 Yrs |  | 12.8 | 23.5 | 23 |  | 54.6 | 100.3 | 100 |
| 8 |  |  | 5+ Yrs |  | 20.3 | 40.7 | 41 |  | 42.8 | 85.7 | 86 |
| 9 |  | R1-R5 | <5 Yrs |  | 26.9 | 53.8 | 53 |  | 43.0 | 86.0 | 85 |
| 10 |  |  | 5+ Yrs |  | 24.9 | 46.2 | 46 |  | 34.7 | 64.4 | 64 |
| 11 |  | M1-M3 | <5 Yrs |  | 11.9 | 11.9 | 12 |  | 10.5 | 10.5 | 10 |
| 12 |  |  | 5+ Yrs |  | 17.9 | 22.6 | 22 |  | 30.3 | 38.4 | 38 |
| 13 | FO | A1-A3 | <5 Yrs |  | 59.2 | 215.6 | 215 |  | 60.9 | 222.1 | 222 |
| 14 |  |  | 5+ Yrs |  | 34.3 | 87.2 | 87 |  | 28.0 | 71.1 | 72 |
| 15 |  | R1-R5 | <5 Yrs |  | 103.8 | 162.8 | 162 |  | 53.3 | 83.6 | 84 |
| 16 |  |  | 5+ Yrs |  | 45.7 | 65.3 | 65 |  | 21.8 | 31.2 | 31 |
| 17 |  | M1-M3 | <5 Yrs |  | 19.3 | 19.3 | 20 |  | 8.1 | 8.1 | 8 |
| 18 |  |  | 5+ Yrs |  | 18.9 | 26.4 | 26 |  | 7.5 | 10.5 | 10 |
|  |  | Total | |  | 575.0 | 1,028.0 | 1,025 |  | 575.0 | 1,064.9 | 1,061 |
|  |  |  |  |  |  |  |  |  |  |  |  |
|  | Objective Function (RelVar) | | |  | 0.0023 |  |  |  | 0.0024 |  |  |
|  | Objective Function (Pct CV) | | |  | 4.82 |  |  |  | 4.87 |  |  |
| 1 Optimized solution from the package.  2 Solution adjusted for sample loss, i.e., optimized solution divided by the eligibility rate times the response rate.  3 Adjusted solution randomly rounded to whole numbers. | | | | | | | | | | | |

1. If the Cycle 5 response rates are less than five percentage points lower than the Cycle 4 values used in the optimization, then there will be a negligible difference in the results.
2. If the difference in the actual and estimated Cycle 5 response rates is approximately 5 percentage points, then the precision of the estimates within the business units will likely fall below the desired CV=0.10.
3. If the actual response rates are more than 5 percentage points lower than the estimated values, then the precision of the business unit estimates will approach a CV of 70%. This is especially true of the SR division estimates because the binding constraint on the sample size as shown in worksheet=“Answer Report 2.”
   1. **Conclusion**

The design team then proceeded to develop a design report around the recommended allocation produced by Excel Solver (Table 7.4). This report included a discussion of the optimization constraints including the need to increase the values for the meaningful detectable differences given the constraint of no more than 600 respondents (i.e., budget). The design team also justified the lowering of the respondent sample size from 600 to 575 by (1) demonstrating the convergence of the optimization system under the reduced sample size, and (2) suggesting that the cost savings could be used on methods to increase participation such as a small incentive.

1. The Excel programmer did not save the various optimization runs as discussed in the VNUV standard operating procedures. The programmer stated that they could be recreated if needed (see Chapter 18 for a discussion of archiving procedures). [↑](#footnote-ref-1)
2. Random numbers from the uniform distribution are generated for each value requiring rounding. If the random number is less than or equal to 0.5, then the integer portion of the value is used as the rounded value. Otherwise, the integer portion plus one is used as the rounded value. [↑](#footnote-ref-2)