Federal Statistical Research Data Center Disclosure Avoidance Methods:

A Handbook for Researchers

VERSION 4.0

The disclosure avoidance methodology in this document comes from guidance by the U.S. Census Bureau's Disclosure Review Board (DRB). This document was released on September 13, 2023. Changes to this guidance will occur, and this guidance is specifically relevant to empirical research output conducted within the Census Bureau firewall. It does not necessarily apply to other types of output or data products. Please consult your FSRDC Administrator or Disclosure Avoidance Officer early and often as you are preparing a disclosure request so they can guide you through the disclosure statistics that are required.

Table of Contents

I. Introduction to Disclosure Avoidance Methods

As a researcher with Special Sworn Status (SSS), you have taken an oath to protect sensitive and confidential data. Products created from internal data are confidential until they have undergone disclosure review and been approved for release.[1] The Data Stewardship Executive Policy Committee (DSEP) oversees data stewardship activities and sets rules and policies that all Census Bureau staff and persons with SSS must follow. The Census Bureau's Disclosure Review Board (DRB) ensures that standard disclosure techniques have been applied and applicable rules have been followed. The DRB also reviews products and project requests to determine if they present additional disclosure risk. Disclosure Avoidance Reviewers (DARs) and Disclosure Avoidance Officers (DAOs) have been trained in disclosure avoidance techniques. A DAO has additional training that gives them *delegated authority* to release output to researchers without bringing output to the DRB for approval. When a DAO determines they cannot use their delegated authority to approve output for any reason (e.g., the current disclosure rules do not directly apply to the requested output, volume of output is too large, etc.), they will bring requested output to the DRB for review.

As part of protecting data, a disclosure avoidance (DA) review process is required before any output may be removed from a secure space (e.g., internal servers or a Federal Statistical Research Data Center). Output includes, but is not limited to, statistical output, notes, programs, graphs, and tables or statements on sign and significance of estimates.

Researchers working at a Federal Statistical Research Data Center (FSRDC) should note that statistical or written output that can be generated outside an FSRDC (e.g., text for a paper, variable definitions, statistics from public data, etc.) *should* be produced outside of the FSRDC. Submitting such items for disclosure review will delay the review process, and any ambiguously sensitive items will be rejected. The process requires researchers to prepare a disclosure request. FSRDC Administrators (henceforth referred to as the RDC Admin) have been trained in disclosure avoidance methods and work with the researcher to prepare a request for disclosure review. Once the RDC Admin determines the requested output is ready to be reviewed by a DAO, they will submit the requested output for review by a DAO in the Center for Enterprise Dissemination-Disclosure Avoidance (CED-DA) group at Census headquarters. The DAO will review the materials and either approve electronic release of requested files, bring the requested output before the DRB, or ask for revisions (rejecting the request if sufficient revisions are not made).

This document explains disclosure avoidance review methods for commonly requested output from empirical researchers working with internal Census Bureau data. As the researcher, it is your responsibility to provide disclosure statistics showing that your output follows these methods. The RDC Admin, DAR, and/or DAO will use these disclosure statistics to conduct their disclosure avoidance review, so it is important to provide all the necessary support files including disclosure statistics, variable and sample definitions, and relationships across disclosure requests. Clear documentation that all disclosure rules have been followed and all disclosure standards are met is essential. **Even if the disclosure risk in practice is low, it is still crucial to properly document that all required privacy-protecting steps have been taken and all disclosure standards are met. If all necessary disclosure statistics and documentation are not provided to the reviewer, the release of the research output will be delayed until it can be confirmed that all disclosure-related items are in order. The disclosure rules and procedures described in this handbook are not all-encompassing. Please consult with your RDC Admin or DAO if a request seems to fall outside the enumerated guidance.**

[1] Throughout this document, any variation of the word "approve" refers strictly to a product being assessed for disclosure risk and confirmed to meet all disclosure requirements for public release. It may also be used interchangeably with variations of the word "clear", "clearance", etc. where "clear" also refers only to meeting disclosure requirements.

There are two main types of disclosure avoidance methods: *legacy methods* (which are still applied) and *modern methods*. Legacy methods include threshold or count rules, rounding, some types of noise addition, and collapsing categories or suppressing cells. For output requests that include economic data, concentration ratios are also used to ensure that statistics are not largely based on a few influential companies. Many FSRDC output requests can be reviewed and approved for release using legacy disclosure avoidance methods. Be advised that the Census Bureau is moving towards modern disclosure avoidance methods that are mathematically proven to maintain privacy protection.

The table of contents should be consulted for specific matters, but the following sections are relevant to nearly all numeric output requests:

II. Samples and Implicit Samples
V.A. Cell Size Thresholds
V.B. Rounding Rules
IV. Volume of Output

- **If working with economic/business data, please also read the following section:**
  V.C. Disclosure Risk from Over-Concentration

- **If generating estimates for substate geographies, please also read the following section:**
  VI. Releases Involving Geographical Areas with Small Populations (GASPs)

II. Samples and Implicit Samples

II.A. Samples

The disclosure statistics described in this document must be provided for all samples and implicit samples. For the purpose of disclosure requests, a *sample* is defined as a set of observations used in an analysis, sometimes specifically referred to as an *analytical sample*. For example, you might define your sample as all firms in a dataset where firm age is five years or greater, while the set of observations actually used in a regression model might include only firms where firm age is five years or greater and where there is non-missing data on all other variables used in the analysis. The latter sample (observations actually used) should be listed as the analytical sample in your clearance request memo, and disclosure statistics should be provided for this sample.

In your requested output files, clearly label which sample is used for each set of estimates. The sample labels (e.g., numbers or letters) in your output files should match the sample labels in your clearance request memo. In your clearance request memo, please refrain from listing samples that do not have any output associated with them, as it will create confusion for the DAO. In your clearance request memo, include samples from a prior release only if they are related to samples in the current requested output. For samples that evolve over different requests, a good practice is to use an expanding naming convention so that you, the RDC Admin, and DAO can see the connection (or lack thereof) between different requests and reduce confusion regarding the need to describe an implicit sample. For example, if request 1111 has Sample 1 and request 1112 uses a subsample of Sample 1, name the sample in request 1112 something like “Sample 2” or “Sample 1a”.

Researchers must provide disclosure statistics for all created samples (and implicit samples). **These disclosure statistics should be calculated at the entity level – the unique firms, persons, or households in each sample or cell.** Counts reported in disclosure statistics should be unweighted. Such statistics include:

- Sample sizes
- Cell counts for categorical/binary variables
- Concentration ratios (for data from economic datasets)

For an analysis conducted above the individual or firm level (e.g., county or industry), disclosure statistics should be calculated for the underlying sample of entities (either individuals or firms). For example, if you estimate a coefficient for an industry-level dummy variable, you should show the disclosure statistics for both the subsample of firms in-sample when dummy = 0 and the subsample of firms in-sample when dummy = 1. If you use something like an industry-level fixed effect and report a mean or percentile for those estimated coefficients, you should show disclosure statistics for whatever underlying sample of firms was included in the calculation. For example, reporting the mean would simply require the disclosure stats for all firms in-sample across all industries (or years, counties, etc.). If you report the pseudo-percentile of such estimates, you should provide the disclosure statistics for the sample of entities whose pseudo-percentile is taken. For example, if you take the pseudo-median of a sample of 125 entities by reporting the average of the middle 11 values, report the disclosure statistics for the full sample of 125 entities. If reporting an estimate for a high-level entity (e.g., the average employment in the highest employment industry), the disclosure stats would pertain to the sample within that entity (e.g., check the firm counts and concentration ratios within the highest employment industry).

II.B. Implicit Samples

Often researchers need to release results based on a main analytical sample and one or more subsamples. Sample sizes and disclosure statistics must be provided for each sample and subsample used to create estimates. In addition, *implicit samples* (sometimes referred to as *complementary samples*) are often created when subsamples are used. Implicit samples are the difference between a larger sample and its subsample. An implicit sample is a sample that can be identified by looking at the differences between explicitly defined populations, samples, subsamples, and geographies. Implicit samples must also be addressed in disclosure review and the usual disclosure criteria apply to implicit samples.

All created implicit samples must be identified in the clearance request memo. Full disclosure statistics, including cell counts for categorical variables, are required for implicit samples except in the following cases:

- **Disclosure statistics for implicit cells are necessary only if the same estimates are requested for release for both the sample and the subsample. For example, if one sample is used for one type of model but its subsample is used for a different model with different variables, then disclosure stats for the implicit sample aren't needed because the estimates produced by the sample and sub-sample are not the same.**
- Sign & Significance Output (see Section III.D. Sign and Significance)

One potentially helpful way to identify implicit samples is to consider whether knowing the sample sizes of two or more samples would yield the true sample size of another sample. If it does, that other sample is an implicit sample. For example, say Sample A contains 1,270 person-year observations and Sample B contains 750 person-year observations. If all 750 observations in Sample B are in Sample A, then an implicit sample of 520 observations appear in Sample A but not Sample B. If some observations in Sample A are not in Sample B, and some observations in Sample B are not in Sample A, we do not have an implicit sample. One may think of such a relationship as "two-way traffic". Other more complex examples are included in Appendix A: Implicit Samples.

Researchers must identify all known implicit samples relevant to their analysis, including those

- Created from different sample definitions from which estimates, samples sizes, or other statistics are being released
- Within a certain release request
- Between a current request and a prior release
- Between the project and other published data, whether from standard publications, other FSRDC projects, or other Census Headquarters projects

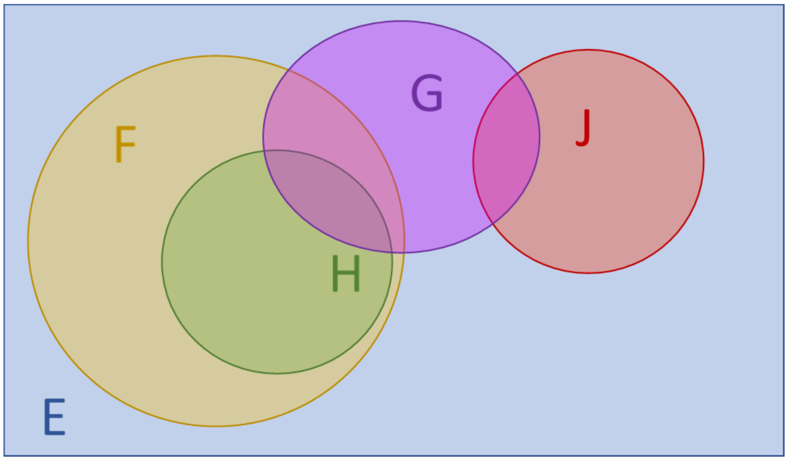
The following examples describe cases of implicit samples:

1. Regression analysis is run on all adults in the American Community Survey (Sample A). The same regression analysis is also run on all homeowners in the ACS (Sample B). There is an implicit sample of non-homeowners (let's call it Sample AB). Your clearance request memo should list this implicit sample (e.g., "Sample AB") and the support files should show person counts for Sample A, Sample B, and Sample AB.
2. Mean employment was released for a sample of establishment-year observations from the Longitudinal Business Database (Sample C) in a previous request (Request 1). Request 1 should have provided firm counts and concentration ratios for Sample C. Request 2 is submitted under the same project. Request 2 asks for mean employment for a subsample of Sample C called Sample D

that consists of establishments with at least 100 workers. Request 2 has an implicit sample (let's call it Sample CD) between Sample C and Sample D consisting of establishments with fewer than 100 workers. The memo for Request 2 should cite Sample CD as an implicit sample, and the researcher needs to provide firm counts and concentration ratios for Sample D and Sample CD. To allow the RDC Admin and DAO to verify that Sample CD is correctly identified as an implicit sample, Request 2 should include the counts and concentration ratios for Sample C from Request 1. In such a scenario, you can simply copy the old disclosure stats from Sample C. New disclosure stats may be required if new analysis is being performed on Sample C (e.g., if a new categorical variable is introduced to the model and those coefficient estimates are being released).

3. Sample E represents a sample of firms in the United States. Sample F restricts Sample E to include only single-unit firms, Sample G restricts Sample E to include only manufacturing firms, Sample H restricts Sample E to include single-unit firms in the 10 most highly populated states, and Sample J restricts Sample E to include multi-unit firms in the 10 most highly populated states. The same regression involving only continuous variables is run separately on each of these samples (i.e., 5 samples and 5 regressions). There are 4 implicit samples of potential relevance here:
    - Sample EF (Sample E – Sample F; i.e., multi-unit firms)
    - Sample EG (Sample E – Sample G; i.e., non-manufacturing firms)
    - Sample EFJ (Sample E – Sample F – Sample J; i.e., multi-unit firms outside the 10 most highly populated states)
    - Sample FH (Sample F – Sample H; i.e., single-unit firms outside the 10 most highly populated states)

    See the diagram below:



Note that any overlapping sample between Sample G and the other subsamples of E is not clearly identified as an implicit sample here.

One can see from these examples that implicit samples can be relatively simple or quite complicated. It is thus extremely important to properly label and describe all samples in the memo and throughout the output and supporting files. See Appendix A: Implicit Samples for more examples and visualizations of potential implicit samples.

Using the earlier example and diagram, supplying firm counts and concentration ratios for Samples E, F, G, H, J, EF, EG, EFJ, and FH should cover all explicit samples and any potentially relevant implicit samples. If researchers are not including disclosure stats for a particular implicit sample, they should clearly note in the supporting documentation why this is not necessary (e.g., different analysis run on each sample, analysis not run on combination of samples).

If special disclosure concerns are present, DAOs have the authority to ask to see full disclosure statistics even when either or both of the above conditions are met. If you have any uncertainty, please consult with your disclosure reviewer.

Researchers should avoid defining samples too early in their projects and avoid publishing results on samples that may change. If a sample changes slightly from one release to the next, it may create a problematic implicit sample that prevents releasing output from the new sample.

III. Types of Output

This section discusses the most common types of files requested for release from the FSRDCs. This is not an all-inclusive list. For each type of output there is a summary of the disclosure statistics required. Details about the required disclosure statistics are found later in this document.

Here is a short summary of possible types of table cells in an output request:

- ➔ Numeric estimates that meet all disclosure requirements can be reported if the relevant rounding rules are applied.
- ➔ Sign and/or statistical significance of regression coefficient estimates may be reported if the unweighted entity count for the sample is sufficiently large (see Section II.D. Sign and Significance).
- ➔ A "D" should be used for suppressed cells that fail disclosure review for any reason (see Section V. When Estimates Do Not Meet Criteria for Release). These suppressions would require complementary suppression if applicable.
- ➔ An "S" can be used for a cell suppressed using user-defined criteria **that does not represent a violation of Census Bureau disclosure avoidance policy**. For example, perhaps some researchers want to suppress cells represented by fewer than 50 individuals. If placing an “S” in a table cell, one wouldn't release specific cell counts, though the suppression threshold used may be acceptable for release if needed (e.g., “S cells have fewer than 50 individuals”). Cell counts should still be provided in the disclosure statistics.
- ➔ A blank table cell is appropriate if a variable drops out of a particular model (e.g., due to multicollinearity), is not included in the model for other reasons, or simply isn't requested for release (but not suppressed for disclosure reasons or user-defined criteria). A value of zero should be reported only if the estimate is actually zero or effectively zero (e.g., through rounding).
- ➔ Estimates explicitly or implicitly conveyed in words/statements can qualify as different types of output. For example, a statement framed solely around sign and/or statistical significance requires less disclosure review compared to language such as “results are similar.”

III.A. Summary Statistics

Standard tabular output, often called *descriptive statistics*, is summary information consisting of, for example, counts, totals, and statistical moments from the distribution (e.g., means). **FSRDC policy is that researchers should limit tabular output to the minimum necessary to describe the sample(s) used in all models and to make pertinent comparisons to the underlying population(s) of interest**. These are the types of tables that usually appear in academic papers.

Researchers often request tables of summary statistics to accompany their model-based output. The following protocols for such tables that must be followed.

1. Test the underlying sample of firms, individuals, or households using our standard disclosure protocols. If the summary statistics are broken into various categories (e.g., sex=0 for male and sex=1 for female), provide the frequency counts for each category so the RDC Admin and DAO can ensure they pass a minimum threshold. For demographic data, examine the number of individuals in the sample. For economic data, examine the number of firms and calculate concentration ratios. In some cases, concentration ratios must also be calculated for each category of categorical variables used in summary statistics. See Sections V.A. Cell Size Thresholds and

V.C. Disclosure Risk from Over-Concentration for more details.

2. All requested output must be rounded according to the rounding rules - see Section V.B. Rounding Rules for more details. Statistics derived from unweighted entity counts, such as proportions, follow special rounding rules described in that section.

3. The FSRDCs typically do not allow true percentiles (including medians), as these generally correspond to actual confidential values.
   - Instead, researchers should calculate a pseudo-percentile. That is, take the mean value from a subset of observations around the percentile – at least five observations on either side, for a total of at least 11 observations for a given quantile.
   - If you compute multiple quantiles using the same dataset, there should be no overlap between the 11 observations for one quantile and the 11 observations for any other quantile.
   - In most instances, if a quantile has 11 unique persons or firms with the same number (e.g., if the median age in a dataset is 33 and there are 11 or more people aged 33 in the dataset) then you may report the quantile.
   - Any reported quantile needs to be rounded according to the rounding rules.
   - The RDC Admin or DAO may determine that reporting a quantile (even when calculating a pseudo-percentile) is a disclosure risk. Such a determination may require additional disclosure statistics, a consultation with the DRB, or rejection of the estimate.

4. Minima and maxima are never released unless there are at least 11 different unique persons or firms sharing the same extremum value for the statistic to be reported. For example, if there are 20 different people earning exactly $150,000 and that is the maximum earnings in the sample, that value could be released if other disclosure requirements are met.

5. If an estimate is derived from percentiles, for example an interquartile range or a difference of quantiles, the published estimate must be calculated using pseudo-percentiles as described above.

6. For disclosure review, you should include full sample disclosure stats for the analytical sample and be able to demonstrate that each reported pseudo-percentile is using a set of at least 11 observations that does not overlap with any other set of observations used to calculate another pseudo-percentile.

7. A STATA .ado file for generating pseudo-percentiles is available. The .ado is called *pseudop*. This program has been vetted by disclosure avoidance specialists. If using the program, please include the .log file as part of support documents so that disclosure reviewers can confirm the source of the pseudo percentiles. If generating your own pseudo percentiles, please include evidence that all rules above were followed.

III.A.1. Output Similar to Public Tables

In some cases, official Census tabulations are already publicly available with similar statistics describing their samples. If the researcher creates a statistic that is roughly the same conceptually as an existing official Census statistic, but without the same disclosure protections, the privacy protections on the official tabulation can be compromised.

To minimize disclosure risks, the DRB requires that researchers seeking to release tabulations using internal microdata assess whether *officially released* Census tabulations would adequately cover their needs in describing model characteristics. The DRB understands that this can be a difficult task, as researchers may be unable to easily determine whether these tabulations already exist.

Researchers should use https://data.census.gov/cedsci/ or other known resources to determine whether tabulations already exist that would meet research needs. The DRB guidance below should be followed to the greatest extent possible, but please understand that it may change over time.

- If it is determined that elements of the research tabular output ***are not*** similar to an official Census tabular release, then the research output remains eligible for DAO approval using delegated authority.
- If it is determined that elements of the research tabular output ***are*** similar to an official Census tabular release, the researcher must do one of the following:
  - ➔ Replace elements with official statistics, and the research product remains eligible for delegated authority.
  - ➔ When noisy inputs are available, create the tabular research output using those noisy inputs. The research product remains eligible for delegated authority.
  - ➔ Provide an explanation and support documentation to justify why the requested research tabular output is sufficiently different from the release and thus does not pose a threat to the privacy protections on the official product. In this case, the research output and this explanation and support documentation must be brought to the DRB for review.

III.B. Model Output

For model output, disclosure analysis focuses on rounding, sample sizes, market concentration (for economic data), and categorical variables. The complexity of the disclosure review will vary depending on the model(s) used. In a regression with dichotomous (0,1) variables, for example, the binary may sometimes take values of 1 for observations associated with a small number of firms. These categories can be disclosive because they identify which firms belong to that specific category.

All disclosure statistics calculations (counts and concentration ratios) listed below should be done at the level of individuals, households, or firms. Even when requested output focuses on different units of analysis such as counties, families, or states, all disclosure statistics should be completed using the individual, household, or firm as the unit of analysis. This may require creating two datasets: one for disclosure and one for analysis. See Sections V.A. Cell Size Thresholds and V.C. Disclosure Risk from Over-Concentration for more details.

You need to make the below calculations **for the observations that actually appear in the model**. Simply using the same sample restrictions is often insufficient. Estimation procedures in SAS or STATA often automatically drop observations during an estimation procedure because of missing values of one of the variables. Make sure those same observations are also excluded from the sample used to calculate disclosure review statistics for that set of regression results or other statistics. Your samples may include different observations from model to model because of these exclusions. If this is the case, **we need a separate set of disclosure review statistics for each sample**. A different N between regressions is a sign that this is occurring. Please make sure the observation count for your disclosure review statistics matches that of the output you are requesting. See Section II. Samples and Implicit Samples for more details.

To summarize disclosure analysis for model output (e.g., regressions):

- The full sample used for each regression must pass the count rules. For economic data, each regression must also pass concentration ratio rules.
- If the model has a continuous variable on the left-hand side and you are reporting indicator (binary) variables on the right-hand side, then make the standard disclosure calculations for all the firms, households, or individuals in each binary category. During disclosure review, such a category may be referred to as a *cell* – the set of observations underlying an estimate. For example, say we have a binary regressor called binaryA and report the coefficient estimate for that variable. One would need to check the cell counts for binaryA = 1 (i.e., the unique people or firms with binaryA = 1) alongside the cell counts for binaryA = 0 (i.e., the unique people or firms with binaryA = 0).
- If the model has a categorical variable on the left-hand side, then make the standard disclosure calculations for all firms, households, or individuals in each category.
- If the model has a categorical variable on the left-hand side and you are reporting indicator (binary) variables on the right-hand side, then make the standard disclosure calculations for all firms, households, or individuals in each category, crossed by the categories on the left-hand side of the model. For example, a binary outcome variable and a binary regressor would create four cells requiring disclosure review (i.e., provide unique person or unique firm counts within each cell).
- If your model has interaction terms constructed from binary variables, the categories for which counts and, if applicable, concentration ratios are required varies depending on what estimates are being released. Below are several examples.
    - ➔ Consider a regression model with binary1, binary2, and interaction12 among the independent variables. If reporting coefficient estimates for all three of these variables, disclosure statistics should be provided for all combinations of binary1 and binary2 (i.e., binary1=0 and binary2=0; binary1=0 and binary2=1; binary1=1 and binary2=0; and binary1=1 and binary2=1). In this case you do ***not*** need to provide disclosure statistics for binary1 and binary2 individually (i.e., binary1=0; binary1=1; binary2=0; binary2=1) unless you are working with economic data for which there could be negative values.
    - ➔ If reporting the coefficient estimates for binary1 and interaction12 but not binary2, provide disclosure statistics for the following categories: binary1=0; binary1=1; interaction12=0; interaction12=1.
    - ➔ If only reporting the coefficient estimate for interaction12, provide disclosure statistics for interaction12=0 and interaction12=1. This relaxes the old guidance which required all category combinations regardless of what estimates were being released.
    - ➔ The above logic extends to cases where three or more binary variables are interacted.
    - ➔ If the interaction term is between, say, one binary indicator and one continuous variable, disclosure statistics for the 0-1 splits are still required. For example, if binary1 and continuous1 are interacted to make interaction11 and the coefficient estimate for interaction11 is released, provide disclosure statistics for binary1=0 and binary1=1 regardless of whether the coefficient estimate for binary1 is being released.
    - ➔ Using a binary outcome variable in the regression model requires the appropriate cross tabs discussed earlier. For example, consider a regression model regressing binaryA on binary1, binary2, and interaction12 and other independent variables. If reporting coefficient estimates for all three of these independent variables, you would need to provide disclosure statistics for all combinations of binaryA, binary1, and binary2 (i.e., binaryA=0 and

binary1=0 and binary2=0; binaryA=1 and binary1=0 and binary2=0; etc.).

- ➔ For a panel event study, provide disclosure statistics to sufficiently cover every 0-1 split pertaining to a reported estimate, as well as any excluded group. For example, consider a dynamic event study with only five time periods, $Y_t = \{0,1\}$ where $t = \{-2, -1, 0, 1, 2\}$, and a binary treatment variable, $D_i = \{0,1\}$, that occurs in time period t=0. It is common to report estimated treatment effects for each lag and lead, while excluding the first lead (t=-1) as a normalization. In this case, the required disclosure statistics include the 0-1 splits for the reported treatment effects and the excluded period. More specifically, provide disclosure statistics (counts and any relevant concentration ratios) for each of the following:

  $\mathbf{1}[D_i Y_{t=-2}] = 1$ and $\mathbf{1}[D_i Y_{t=-2}] = 0$

  $\mathbf{1}[D_i Y_{t=-1}] = 1$ and $\mathbf{1}[D_i Y_{t=-1}] = 0$

  $\mathbf{1}[D_i Y_{t=0}] = 1$ and $\mathbf{1}[D_i Y_{t=0}] = 0$

  $\mathbf{1}[D_i Y_{t=1}] = 1$ and $\mathbf{1}[D_i Y_{t=1}] = 0$

  $\mathbf{1}[D_i Y_{t=2}] = 1$ and $\mathbf{1}[D_i Y_{t=2}] = 0$

  Note, if the event study uses a binary outcome variable, then disclosure stats for cross tabs of the outcome variable with each of the ten binary interaction terms listed above would be required, i.e., 20 sets of disclosure stats.

- You need to provide disclosure statistics for only binary variables that you report. For example, if you use dummies as control variables or fixed effects and do not report the associated estimates, then you do not need to provide disclosure statistics for such cells.

- More generally, if you have a discrete variable that is being used as a categorical variable (e.g., mean earnings by number of children, or a series of dummy indicators for family size), provide disclosure statistics for each category and any relevant cross tabs. If the discrete variable is being used as a continuous measure (e.g., using “number of children” itself as a regressor or outcome variable), the researcher should confirm that the variable does not reduce to a binary indicator in their samples and/or analysis. For example, say that “number of children” ranges from 0 to 12 in the full sample, but for a certain subsample there are only 0 and 1 values. Those cell counts (and, if applicable, concentration ratios) would need to pass disclosure requirements even if the variable is being treated as continuous in the model.

- If the model is being run on multiple samples that create an implicit sample, all counts and concentration ratios must be documented for the implicit sample as well. See Section II.B. Implicit Samples for more details.

- Relatedly, if alternate forms of the same categorical variable are being run on an identical model specification and analytical sample, all counts and concentration ratios must be provided for any implicit cell created. For example, regressing wages on a binary indicator for college where college is defined as “some college or higher”, and then running the same regression but with college defined as “college degree or higher” would create an implicit cell where college is a binary variable for “some college”. In this case, you would need to report counts and any necessary concentration ratios for the binary variable “some college”.

- For an instrumental variables regression, the necessary disclosure statistics depend on whether the first stage is being reported. For example, consider a simple IV design where the outcome variable Y is continuous, and an endogenous binary variable X is being instrumented by continuous variable Z. If only second-stage results are being reported, i.e., the regression of Y on predicted variable $\hat{X}$, then categorical splits are not required since $\hat{X}$ is continuous – only disclosure statistics for the overall sample are necessary. However, if the first-stage results are being reported, i.e., the regression of X on Z, then three sets of disclosure stats are required: X=0, X=1, and the overall sample. Note, if Z is also a binary variable, then disclosure statistics for the cross tabs of X and Z would be required when reporting the first-stage results; i.e., X=0 and Z=0, X=1 and Z=0, X=0 and Z=1, X=1 and Z=1.

- All coefficients must be rounded to four significant digits. See Section V.B. Rounding Rules for more details.

III.C. Graphical Output

Data can often be presented more effectively as a graphic than as a table of numbers. It is your responsibility as a researcher to minimize the amount and precision of data leaving the Census Bureau to reduce the chance of an unlawful disclosure.

Graphical output is subject to the same general disclosure standards as non-graphical output: **any information that is based on a small sample or a highly concentrated cell will not be released**. If the data used for the graph/figure will be released, the data must also pass all disclosure review rules. Once any underlying data or estimates pass disclosure review and are released, you may create any graphs or figures using that released data. The guidance in this section is for graphical output where the underlying data or estimates are not being requested for release.

With figures like kernel density plots as a notable exception, it is recommended that researchers release the underlying data for most types of figures as a table instead. Such estimates still need to follow all disclosure rules (e.g., rounding). Using the estimates approved for release, the researcher could then reformat the figures as needed or make as many figures as desired outside of the FSRDC without requiring additional disclosure review.

The preferred method for creating graphical output is to round the underlying data (see Section V.B. Rounding Rules) and produce the graph or figure with the rounded data or estimates. If the underlying data or estimates cannot be shown to be rounded to the appropriate level, the graph needs to have a low enough resolution such that all rounding requirements are met.[2] If the underlying estimates pass all necessary disclosure rules (including rounding), the resolution of the figures is inconsequential.

Regardless of whether the underlying data are also requested for release, you need to provide the appropriate disclosure statistics for all graphs and figures. Furthermore, graphical output counts toward volume of output (see Section IV. Volume of Output).

All figures must be produced using one of the following approved formats:

- .png files
- .jpeg files
- .tif files

[2] Rasterized (i.e., simple bitmap) images should be rendered at the intended publication size (e.g., 3" x 5") and at or below 300 dots per inch (DPI). In the 3" x 5" example, this effectively bins data to no larger than 900 x 1500 bins. Furthermore, the scaling of the graph is relevant. If the x-axis and y-axis scaling include multiple significant digits, the DPI may need adjusting to account for the scaling. For example, having 0.01-unit intervals on an axis would require lower DPI than having 0.1-unit intervals because the level of detail of the former is greater. Knowing a point falls between 0.32 and 0.33 provides two significant digits of information whereas knowing a point falls between 0.3 and 0.4 provides only one significant digit of precision. In short, if the resolution meets these standards, the underlying data presented in the graphs or figures would effectively meet the rounding requirements.

III.C.1. Kernel Density Plots

You may wish to produce supporting figures or graphs to describe the data used in your projects. One way to do this is to produce a histogram, which graphically displays the univariate distribution of such data. Instead of histograms of univariate distributions, which give counts of numbers of observations within certain classes, researchers should produce kernel densities, which are essentially smoothed versions of histograms. In estimating kernel densities, the researcher should choose bandwidth values that do not obviously suggest the presence of individual observations. Furthermore, the bandwidth value itself should not be released.

For multivariate distributions, scatterplots of data on individual observations are generally not permitted. For example, a scatterplot of value of shipments versus employment for individual establishments/firms would not be allowed. Instead, bivariate kernel densities, which do not show the individual observations, should be produced.

These considerations also apply to data that are not in the scale of the observation, e.g., productivity measures, which are ratios and not tightly related to size. Researchers should think of alternative ways to show these relationships, e.g., bivariate kernel densities.

The DRB has instituted the following protection measures for releasing a kernel density plot:

1. Test the underlying sample of firms/households/individuals using standard disclosure protocols. Report unique entity counts for each bin. Each bin needs to have at least 3 unique firms/individuals/households). This is required even for kernels with infinite support (e.g., Gaussian). Bin sizes can be generated as if a histogram were being produced using the bins from the kernel density by using the plotted x axis points as the midpoints of the bins. An alternative is to use the bandwidth and the given density point estimate (e.g., use STATA's generate command and take +/- r(bwidth) for each x-axis point). Note that this method will need to account for overlapping bins for non-Gaussian kernels and may provide nonsensical results for a Gaussian kernel. For economic data, in addition to reporting the number of firms within each bin, you may need to provide concentration ratios. There are 2 possibilities:

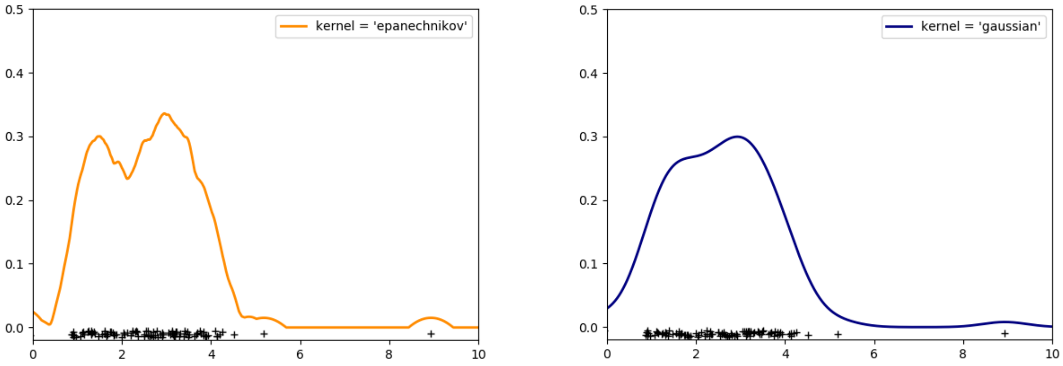
    - A kernel with infinite support requires only full sample disclosure stats, i.e., CRs for the full sample used to generate the kernel density plot.
    - If using a kernel with finite support, such as the Epanechnikov kernel – which is the default in STATA – check the full disclosure stats within each bin.

2. Cut off 5% from each tail. This eliminates the possibility of releasing information on extreme values. The 5% can be cut off before the kernel density estimator (KDE) is run or after the KDE is run. Removing the tails is not required when the estimates for which the distribution is being generated would pass disclosure on their own. For example, if you run a series of simulations and produce a KD plot of regression coefficients, you do not need to remove the 5% tails. All relevant disclosure statistics are required in such cases.

3. Limit the detail of the scales on both axes. If values must be included, researchers should use broad rounded numbers, not the values that are automatically generated by various software programs. In addition, you must suppress the bandwidth used in producing the kernel density. Some programs include a bandwidth value on the plot as part of a key or legend. A data intruder can subtract this bandwidth value from the maximum value labeled in the plot and closely estimate a true actual maximum value. Assuming no numbers are released, it is acceptable to include a note saying that the default bandwidth from your statistical software was used.

4. Rounding the underlying data to four significant digits before calculating a kernel density is strongly recommended. There is a program called *rounddig.ado* available to researchers. You can round the underlying data to 4 significant digits by placing "rounddig 4 <variable name>" in your code. If you have significant concerns about degradation of data quality using this approach, you may use the raw data and limit both the detail of the axes and the resolution of the figure. The disclosure review in such cases will be more intensive and likely take more time.

Plots of the kernel density generally have the variable of interest along the X-axis and the density along the Y-axis. These rules are intended to protect an individual's (or firm's) value regarding that variable of interest. In particular, the bin count rule (#1 above) ensures that the data do not reveal the presence of a single entity in a sparsely populated area of the plot. Even kernels with infinite support do not always protect this information, so the bin count rule is required for all KD plots. Consider the following example where both the Epanechnikov and Gaussian kernel reveal that there is a single entity with a value of 9:



Fixing the problem presented in these graphs requires either dropping the questionable entity before running the KDE or increasing the bandwidth size so that requirement (1) is satisfied.

There are a few additional notes to consider with KD plots:

1. If you are generating a KD plot with more than one sample, and if this creates an implicit sample, then you **need to check implicit sample disclosure statistics only if the bins are the same**. In that case, you would need disclosure statistics for the implicit samples per bin.
2. If you are using observations at a higher level than person or firm (e.g., industry or part-state geographies) **and those data are external**, then check those entity counts as well. For example, for external county data linked to an internal sample of firms, check bin counts for both firms and counties.

III.D. Sign and Significance[3]

As outlined later in Section IV. Volume of Output, the threat of disclosure risk from advanced attacks, like a database reconstruction attack, increases considerably as more information is released. One way to minimize the volume of output is to report only the sign and significance of an estimate rather than its value and associated variance. Reporting sign and significance (sometimes abbreviated "S&S") can be an

[3] This was previously referred to in the FSRDCs as "Qualitative Output". To clarify what qualifies as qualitative output and to differentiate from the DRB's qualitative output terminology referring to information products based on interviews or content analysis, the FSRDCs now use the 'sign and significance' terminology.

alternative to reporting numeric estimates as robustness checks for journal submissions, as requested output to be used at presentations for feedback on your research, and even for publications. In addition, the disclosure review process for sign and significance tables or statements is typically faster than a numerical output disclosure review.

Sign and significance output generally does not count towards volume of output. The DAO will use their judgment to determine whether a large amount of sign and significance, or a large amount in combination with numeric estimates, can be approved using delegated authority or whether the requested output needs to be reviewed by the DRB.

Disclosure statistics required for sign and significance output can vary depending on the samples and datasets used. If only sign and significance are being reported, then typically only overall explicit sample sizes are required. If some numeric values are being reported while others are replaced with sign and significance, then all disclosure statistics are required. When any numeric estimates are being released, you must track relevant implicit samples and provide appropriate disclosure statistics. **Note that any sign & significance output that could pertain to Geographical Areas with Small Populations (GASPs) would still require population analysis** – see Section VI. Releases Involving Geographical Areas with Small Populations (GASPs).

Consider the following example. Suppose the same analysis is run on two different samples, A and B, and produces numeric estimates. Then 100 signs and significances are generated for samples A, B, C, and D. The following disclosure statistics would be required:

1. The full disclosure statistics for A, B, and any implicit samples created by A and B.
2. The unique person or firm counts in samples C and D.

In this scenario one would NOT need to worry about any implicit samples involving C or D (i.e., implicit samples between A and C, A and D, B and C, B and D, or C and D) since numeric estimates aren't being released for C and D. If numeric estimates for C and/or D are requested later, full disclosure statistics and implicit sample considerations would become necessary for disclosure review.

If you follow all guidelines in this section, the benefits of sign and significance output include expedited disclosure review, fewer disclosure statistics required, and no volume of output considerations (within reason). Your RDC Admin and DAO may determine on a case-by-case basis that additional disclosure statistics are required when sign and significance are reported. For any prose or result summaries not framed solely around sign and/or statistical significance (e.g., “the results are similar,” “the coefficient is large”), you must provide all appropriate disclosure statistics.

Sign and significance output can be released in a few ways. The following examples are not exhaustive, but the simplest scenarios all involve only reporting the sign and/or statistical significance of the requested estimates:

1. All estimates in a table can be replaced by sign and significance.

	Sales
Firm Age	+**
Firm Size	+*
Number of Observations	10,000

You can report the number of observations, but that number needs to be rounded and will count as one estimate towards determining the volume of output.

2. Some estimates in a table can be replaced with sign and significance while the value associated with the main variable(s) of interest can remain.

	Sales
Firm Age	+**
Firm Size	2.34 (.567)
Number of Observations	10,000

Here, two estimates (coefficient and standard error for Firm Size and Number of Observations) count towards the volume of output but the sign and significance of Firm Age does not.

3. Sign and significance output may also be written in sentences. For example, "The coefficient of Firm Age is positive and statistically significant when controlling for Firm Size" would qualify for sign and significance review.
4. Something like "The coefficient estimate for type X is not statistically different from the coefficient for type Z" or "the difference between the mean earnings for group W and group Z is statistically different from zero" would be eligible. If no numeric estimates are being released, then the statement is fully S&S output, meaning that only the explicit sample size (entity count) is required and there is no volume of output. If one number is being released but the other isn't, the full disclosure stats are required for the numeric estimate and the unreleased estimate would still count toward volume of output since its magnitude is being established to a greater degree than the standard S&S request. If both numbers are being released, a statement on sign and/or statistical significance of the difference (or lack thereof) between the estimates would not add to the volume of output.
5. If alternate language (e.g., "results are comparable," "estimates are similar," "the effect is large") is used, such language will not qualify for sign & significance review. The estimates would be treated as if they were being requested for release in the regular fashion and would require the usual disclosure statistics for the disclosure review. Researchers should then provide in the support materials the referenced tables/estimates, the table shells, or describe such tables in the support documentation (e.g., "same table as Table 2 except adds in control variable X").

III.E. Programs

Most programs created by researchers may be released after being reviewed by a DAR or DAO. The initial review consists of determining whether the program/code is "in-scope" for disclosure review or "out-of-scope". If the reviewer can determine that the program/code is out-of-scope, then the program is releasable as is and requires no disclosure approval number. Otherwise, revisions (e.g., redaction) may be required to render the program/code out-of-scope, or a full disclosure review may be necessary.

Neither the code itself nor the comments may reveal any information about the underlying data, regardless of whether the information revealed falls under usual definitions of PII. All such information must be redacted or go through the normal output review process. If a special case such as an outlier requires special treatment, researchers should take extra care that the part of the code referring to the special case does not reveal anything about a small number of records. Code or programs that explicitly include or implicitly describe counts, sample sizes, estimates, etc. would be in-scope for full disclosure review. Code comments can be problematic if they contain references to the sample, e.g. "Only a few observations were dropped here", or "No values in sample A changed in this code block". It is best to ensure that comments are kept to a minimum to avoid such issues.

If a program creates a subset of the data (e.g., by county or NAICS code), the researcher should confirm that the subset criteria was implemented without referencing the internal dataset. For example, if a researcher uses a survey that samples only some counties, subsetting the data to refer only to particular counties could potentially reveal which counties are in the sample. Releasing programs showing such a list would not be permissible, unless the list was made before knowing which counties are in sample and thus may include in-sample and out-of-sample counties. Similarly, if a program includes a variable for whether a household has over or under the median income, the median income may not be hard-coded into the program.

Programs that are fit for public release cannot include identifier values (e.g., PIK = 12345 or EIN = 54321). Furthermore, programs released publicly cannot include non-public configuration information about Census Bureau computing systems. Programs released publicly cannot include a "James Bond ID" (JBID) for a Census Bureau data user, including the researcher. All pathnames in the code should also only contain relative paths. Full pathnames and JBIDs (if applicable) should be removed from the program.

Examples of unacceptable and acceptable pathnames:

Unacceptable

```
\\int123-internal\user\bond007\my_project\programs\my_code
```

This portion contains the full pathname and a JBID.

Acceptable

```
\my_project\programs\my_code\filename
```

To reiterate: programs requested for public release should not explicitly contain or implicitly reference any sensitive or potentially sensitive information. Put another way, programs/code requested for public release ideally could have been written outside the firewall. Anything ambiguous will be flagged by the code reviewer with the release of the program/code delayed accordingly.

With more journals requiring researchers to submit their code along with their research findings, disclosure reviewers have seen a substantial increase in the number and size of code requests submitted for release. Researchers can shorten disclosure review time by modularizing code – creating macros or similar routines to do repetitive tasks rather than repeating nearly identical code several times in the program. This approach is generally considered good programming practice, and it makes the code quicker to review.

IV. Volume of Output

The threat of disclosure risk from advanced attacks, like a database reconstruction attack, goes up considerably as more output is released from the same sample of microdata. Therefore, items with a large amount of output require DRB approval. As a researcher you should always ask yourself:

- "Do I need so many cells to answer the research question?"
- "Am I being asked for so many cells for external review?"
- "Can I collapse table categories or geography?"
- "Can I consider other ways to present my findings?"
- "Am I accounting for implicit samples in this and across prior releases as more cells might be derived by subtraction?"
- "Am I constraining possible revisions that may create implicit samples with the current output?"

Researchers should only request output for which they have a clear need. In practice, it can be difficult to determine essential estimates, but there is a limit on the total number of estimates that can be released (exactly what counts as an estimate is discussed later in this section). Furthermore, increasing the number of estimates in each request reduces the likelihood of approval. This is particularly true as researchers produce additional requests in the same line of research.

IV.A. What counts as an estimate?

First, it is useful to describe how to tally amount of output. Any of the following could constitute a single "estimate" for the purpose of disclosure review:

- A single point estimate and its corresponding measure(s) of variance
- A test statistic such as a p-value, F statistic, etc.
- A cell in a variance/covariance matrix
- Number of observations or entity count for a sample

The sample size or number of observations only counts toward volume of output once per sample. For example, releasing the information point that Sample A has 100,000 firms would only count as one estimate even if "N of Sample A" is included in multiple tables. If an estimate requested for release could be derived from other released estimates, that estimate is not included in the volume of output count (e.g., confidence intervals are not included in volume of output counts if standard errors, sample size, and coefficient estimates are also requested for a regression model). However, releasing additional standard errors (e.g., clustering at a different level) adds to volume of output, with each additional standard error counting as a half estimate. So, if you request 20 coefficient estimates and standard error pairs and then cluster those standard errors at a different level, the re-clustered standard errors count as 10 estimates to the volume of output.

Graphical output and figures count toward both volume of output thresholds described below. Every grid point representing a data point counts towards the number of estimates. For example, a kernel density plot with 40 grid points (data points) counts as 40 estimates. Figures may be viewed more favorably from a disclosure risk perspective if the resolution of the figure reduces the precision of the estimates.

Across related requests, the sum of estimates from numeric and graphical output will be the metric for whether a request is eligible for delegated authority or requiring DRB review. Examples of volume of output calculations are included in Appendix B: Volume of Output Examples of this handbook.

IV.B. Volume of output tracking

Volume of output is cumulative across requests that use the same research sample for analysis. Researchers should thus try to plan ahead on how much volume of output they could be requesting on a given sample or set of related samples – especially if the researchers expect to be submitting their work to conferences, journals, etc. and returning for additional estimates based on feedback received. Researchers should also be careful about requesting numeric estimates that may not be final, as multiple "versions" of an estimate will count separately toward amount of output. For example, say that you release an estimate and later re-generate the same estimate after correcting a coding error or using different weights. Releasing the second estimate would still count toward volume of output. In some cases, updating previously released numeric estimates could present additional disclosure risk unrelated to volume of output.

If the first request consists of several thousand estimates, future requests derived from the same sample(s) would most likely require DRB review. Researchers should keep track of the samples used across requests as well as what types of analysis were performed on which samples. Further, researchers are responsible for describing samples' relationship with past releases from the same project. Failure to do so will delay the disclosure review. It can of course be difficult if not impossible to know all previously released output that used a particular data sample. The expectation is that the researcher can at a minimum keep track of the samples used in their own research within and across FSRDC projects. When possible, researchers should be aware of other research being performed on related data samples under the same project.

The DRB has set two limits on the number of estimates that can be released by a DAO without DRB review. These two limits are not exhaustive. An RDC Admin or DAO may recommend that due to a request's particular features, a request should be brought to the DRB over volume of output concerns. In particular, the DRB will likely be asked to make the final approval if a large number of estimates are being generated for a small sample of entities – even if neither volume of output limit is technically violated.

First, the DRB has set a nominal limit of 5,000 estimates for delegated authority. This cap is cumulative *across* related requests involving the *same or directly related* analytical samples. A DAO can exercise delegated authority on up to 5,000 estimates from the same analytical sample; however, DRB review is automatically required once the cumulative volume of output exceeds 5,000 estimates.

Second, the DRB has also set a relative volume of output threshold. Output requires DRB approval if the ratio of the *unweighted sample size* (count of unique firms, people, or households) for the "main" research sample(s) to the *total number of estimates* derived from that sample(s) is less than 30:1.

Below are some hypothetical examples of applying the 30:1 rule in practice:

- If Samples 2 and 3 are subsets of Sample 1, the 30:1 ratio could be the unique entity count in Sample 1 divided by the total number of estimates requested for any analysis using Samples 1, 2, and 3, or it may be based on each sample separately.
    - ➔ Suppose the analysis on Sample 1 involves multi-unit firms, while Samples 2 and 3 are used for a supplemental analysis on multi-unit firms in the manufacturing industry and services industry, respectively. In this case, the entity count used to calculate the 30:1 ratio would be the number of unique multi-unit firms in Sample 1, and the number of estimates would be the sum of all estimates requested for release using Sample 1, 2, or 3.
    - ➔ Suppose Sample 1 is identified as a distinct research sample (i.e., not the entire survey data) but is ***not*** used for any analysis, then calculating the relative volume of output ratio would depend on the relationships among samples. For example, if Sample 2 and Sample 3 are disjoint (e.g., men and women in the survey), then separate 30:1 ratio calculations would

be applied. However, if there is overlap between Samples 2 and 3, the 30:1 ratio rule would typically treat Sample 1 as the "main" sample for volume of output calculations.

➔ Suppose Sample 1 is from dataset A and Sample 2, a subset of Sample 1, merges on information from dataset B, then Sample 1 and Sample 2 should be treated separately for 30:1 ratio rule considerations.

- If a request is comprised of multiple samples with no clear "main" research sample, the 30:1 ratio would be applied separately for each high-level sample. For example, suppose a request involves analysis on Sample D comprised of firms in the healthcare sector (along with assorted subsamples) and Sample E comprised of firms with 50 or more employees (along with assorted subsamples). In this case, Sample D and E overlap but neither is a subset of the other. So, a 30:1 ratio would be calculated using all estimates derived directly or indirectly from Sample D, and another separate 30:1 ratio would be calculated for Sample E using its associated estimates.
- For panel data where the output is not reported by year, the unique entity is the person, household, or firm, not crossed by year. For example, if your sample consists of 1,000 unweighted individuals and individuals respond once a year for five years, the sample size used to calculate the 30:1 ratio would be 1,000. If, instead, estimates were reported by year, the 30:1 ratio would be calculated using entity-year observations.
- If the overarching research sample is very large but a relatively high amount of output is being generated on a very small subset, the output may be taken to the DRB. For example, if one table of summary statistics is generated on Sample 1 with 500,000 unique individuals, but there are 100 regression estimates generated from a subsample of Sample 1 consisting of only 500 individuals, such a request would require DRB approval.

The above list of examples is far from exhaustive. There are many possible scenarios, and the volume of output rules can be difficult to enforce in practice. Researchers working with small analytical samples and/or requesting a large amount of output should have their request ready for review several weeks before any deadlines to permit enough time for the DAO to prepare the request for DRB review. Again, the DAO always has the authority to take any ambiguous or otherwise concerning cases to the DRB for final review/approval even if the volume of output is seemingly eligible for delegated authority.

Attempts to circumvent these rules by splitting up output requests, slightly altering samples, or adding larger samples to the request, etc. may result in rejected output. At the very least, such requests would be treated as a single collection and taken before the DRB for review. Going to the DRB does not imply a request will be rejected, but under current guidelines a DAO cannot use their delegated authority to approve requested output exceeding the limits. If the RDC Admin or DAO determines that researchers on a project are trying to "game the system", they may recommend those requests be rejected outright.

Researchers should consider the following ways to limit their project's volume of output:

- Consider not reporting regression coefficients on less important variables (e.g., geographic indicators, sector-specific dummy variables, or fixed effects).
- Consider aggregating dummy variables (e.g., reporting estimates by sector instead of subsector).
- Instead of reporting numeric parameter estimates for robustness checks, or any other less important regressions, consider reporting the sign of the estimates and their significance levels.

V. Legacy Methods

Legacy methods are the traditional techniques applied to any data releases from the Census Bureau. While some releases have begun to use modern methods (see Section X. Modern Methods), legacy methods are still used for many empirical output requests. The basic idea to keep in mind when applying disclosure rules is that each statistic, coefficient, number, etc. to be released has an underlying sample of original person, household, or firm observations in the microdata. That underlying sample must pass all relevant disclosure rules, as documented in supporting statistics. Researchers need to prepare disclosure statistics to show that their requested output passes disclosure avoidance legacy methods for all requested output, for each sample requested, and for each implicit sample created. Below are the disclosure avoidance legacy methods.

V.A. Cell Size Thresholds

The Census Bureau requires a minimum unweighted count for each cell (i.e., the observations explicitly or implicitly described by the reported estimate). Counts of zero are not considered a disclosure risk because you cannot learn any more about an establishment, individual, or household where a cell size is zero. However, very small counts are a disclosure risk. **At minimum, unweighted cell size must be at least three.** The Census Bureau also requires microdata and tabular data to meet certain thresholds. See References for more information.

For disclosure review, unweighted cell size counts must be provided for all samples, subsamples (including counts for all categories of dummy and categorical variables), and implicit samples.

For Title 26 counts and estimates from Internal Revenue Service (IRS) data and commingled data (from the Census Bureau and the IRS), we enforce the following thresholds based on IRS requirements:

<u>For establishment data:</u>

- At least 3 companies (firms) for national estimates.[4]

- At least 10 companies for state-level estimates.

- At least 20 companies for substate-level estimates, except for zip codes.

- At least 100 companies for ZIP code-level estimates.

<u>For housing unit data:</u>

- At least 3 housing units for national estimates.

- At least 10 housing units for state-level estimates.

- At least 20 housing units for substate-level estimates, except for zip codes.

- At least 100 housing units for ZIP code- level estimates.

[4] A collection of multiple states qualifies as national level output in terms of Pub 1075, as long as the output is not broken down by state. Consequently, estimates derived from a pooled sample of multiple states only requires a minimum cell size of three.

V.B. Rounding Rules

Appropriate rounding methods are generally required for all statistical output, such as summary statistics, tabulations, and model-based estimates, including but not limited to coefficients and standard errors. Released counts for highly aggregated entities (e.g., industry or county) do not need to be rounded if they are public knowledge. If they are internally derived (e.g., number of counties present within an internal sample), counts must be rounded according to entity count rules. Rounding is a legacy method that reduces information in a cell. Rounding rules applied to unweighted counts, weighted estimates, and model-based output are occasionally revised to better protect the data. Note: if you use modern disclosure avoidance methods (e.g., noise injection, formal or non-formal privacy methods), you are not required to round your requested output.

The current rounding rules are different for weighted vs. unweighted counts.

- An ***unweighted count*** is the actual number of observations used in a statistical analysis.
- ***Weighted counts*** are unweighted counts projected to a larger population. Final weights are applied at the individual person, household, or establishment level, then aggregated and rounded.

V.B.1. Rounding for statistical output and weighted counts

All publicly released statistical output and weighted estimates must be rounded to four significant digits, base 10. Four significant digits is defined as

x.yyy multiplied by 10^{nnn}, where 1 <=x <=9, 0 <= yyy <=999, and nnn is the exponent.

These numbers all have four significant digits:

Numbers with four significant digits	**Scientific notation**
1,234,000.	(1.234E+006)
1,234.	(1.234E+003)
1.234	(1.234E+000)
0.0001234	(1.234E-004)

The above guidance applies when statistical output and weighted estimates are reported in thousands, millions, or any other unit. Trailing zeros after a decimal point and beyond four significant digits (e.g., the zeros in 1,234.00) are not allowed and will not pass disclosure review.

V.B.2. Significant digits

Significant digits are defined as the number of digits in a numeric string that starts with the first non-zero digit. Every digit after the last significant digit can be a zero except if those zeros are after a decimal place.

Zeroes can be part of the set of significant digits beyond the first significant digit. For example, if you are told that the number 1,000,000 was reported to 4 significant digits, then you know the unrounded number lies in the range [999500, 1000500). Alternatively, if you were told 1,000,000 was reported with one significant digit, then you know the unrounded number falls in the much broader range of [500000, 1500000).

All weighted and unweighted summary statistics, tabulations, model-based estimates, and all other estimates (excluding unweighted counts) must be rounded to four significant digits as described above. This includes, but is not limited to, the following:

- Means
- Model coefficients
- Standard deviations/standard errors
- Variances, covariances, and correlations
- Test statistics
- Weighted estimates

In most cases, this rounding method will not have a substantive effect on the ability of information product originators to make correct inferences with their data. However, rounding may prove a hindrance in special cases, e.g., in simulation studies. If you have a demonstrated analytical need for an exception, work with your RDC Admin or DAO to request an exemption from the DRB.

V.B.3. Rounding for unweighted counts

Researchers often seek to publicly release unweighted counts for each cell in an analysis. These counts can be disclosive and are discouraged. Instead, a suggested course of action is to compute the estimate’s standard error or margin of error for each cell (e.g., the 90% confidence interval that appears in most official Census Bureau publications). If the researcher reports the weighted estimate and its associated standard error or margin of error, then there will be no need to report the number of observations in the cell for external release.

To assist in disclosure review, please provide unweighted and unrounded counts in your support files. This allows the reviewers to check sample relationships and ensure output files have been rounded correctly.

The rounding rule for unweighted counts is as follows:

- If N is less than 15, report N < 15
- If N is between 15 and 99, round to the nearest 10
- If N is between 100-999, round to the nearest 50
- If N is between 1,000-9,999, round to the nearest 100
- If N is between 10,000-99,999, round to the nearest 500
- If N is between 100,000-999,999, round to the nearest 1,000
- If N is 1,000,000 or more, round to four significant digits as described earlier.

For unweighted counts where N < 15, you will need to additionally suppress associated statistics such as standard errors, percent coefficients of variation, rates, and ratios. Counts for observations derived by, or closely related to, entities such as person-years or firm-years are also subject to the unweighted rounding rule.

Administrative record files and the Census Bureau’s internal frames generally do not have weights associated with each observation. When using administrative record files, counts of entities (households, persons, jobs, or establishments) are subject to the above unweighted rounding rule. This includes administrative files such as the NUMIDENT, LBD, and LEHD Infrastructure (excluding QWI).

All reported numbers of observations (entity counts in the dataset used for the estimation) must be rounded according to these rules, even for large Ns. This rule is designed to limit difficult-to-detect disclosure of sliver populations. An example of a sliver population is a small industry that is part of a very large sector.

This rounding rule is not intended to undermine scientific validity. Most researchers report the number of observations in summary tables. If you need a full-precision unweighted count, then this is inherently problematic. If you are using the count to indicate the size of the sample, then you need to follow the unweighted rounding rules.

Other integers related to observation counts must also be rounded. These include error degrees of freedom, and degrees of freedom associated with entity-level fixed effects (person effects, household effects, firm effects, establishment effects, job effects). Degrees of freedom associated with model effects, including high-dimensional effects that have standard categorizations (block effects, tract effects, county effects, state effects, NAICS effects, SIC effects, etc.) should be rounded according to the four significant digits rule, not this unweighted rounding rule.

Degrees of freedom that are directly derived from previously reported rounded numbers need to be consistent with those previously reported rounded numbers, even if they are not in the same information product. For example, if the number of categories of a variable is rounded to 30, then the degrees of freedom in a fixed effects model related to that variable would be reported as 29.

V.B.4. Rounding for ratios or proportions

Derived measures, such as ratios or proportions that are based on unweighted entity counts and totals, should be calculated using rounded numerators and denominators. Proportions and ratios can be calculated using unrounded numerators and/or denominators if you limit the precision of the resulting ratio. For thresholds based on an unweighted denominator (D) and an unrounded unweighted proportion (P):

- If 15 <= D <= 100 then P should be rounded to 1 significant digit
- Else if D <= 1,000 then P should be rounded to no more than 2 significant digits
- Else if D <= 10,000 then P should be rounded to no more than 3 significant digits
- Else if D > 10,000 then P should be rounded to no more than 4 significant digits.

The “D” is based on the unweighted denominator rounded according to unweighted entity count rounding rules described earlier. When using this threshold, any reported numerators and denominators must likewise be rounded according to the unweighted entity count rounding rules. Unweighted numerators or denominators less than fifteen will be recorded as N < 15. In the earlier example, the reported proportion would be 0.594 since D is greater than 1,000 but less than 10,000. For panel data, the “D” above pertains to the rounded number of (unweighted) observations. For example, if there are 100,000 observations in the denominator but only 500 firms, the proportion P may be rounded to 4 significant digits.

Another rounding option for unweighted proportions is to first round both the numerator and denominator according to the unweighted entity count rules (see Section V.B.3. Rounding for unweighted counts) and then report the quotient to 4 significant digits. In situations where the two rounding rules do not align (e.g., when the denominator-based method would allow 3 significant digits, but the other method would require rounding to 4 significant digits), researchers should indicate which method they used.

V.C. Disclosure Risk from Over-Concentration

In economic data, the distributions of variables are often highly skewed so that a few firms (even one) could account for most of the value in a cell, i.e., the cell could be highly concentrated. The Census Bureau does not allow the release of output that comes "too close" to revealing any individual firm's value or presence in the sample. Two rules for identifying unallowable cells are the *p% rule* and the *(n,k) rule*. These rules are based on the idea that the other firms contributing to the cell total (the firm's competitors) are the ones best positioned to determine the values contributed by the other firms represented in the cell. These rules acknowledge that the other firms contributing to the cell total (the firm's competitors) are the ones best positioned to determine the values contributed by the other firms represented in the cell.

Under the p% rule, which is used for Census Bureau data from 1992 and later, a cell is sensitive if the cell total with the two largest firms removed is greater than p% of the value for the largest firm. The value of the parameter p is highly sensitive. Under the (n,k) rule, which is used for Census Bureau data before 1992, a cell is sensitive if the largest n members of the cell contribute more than k% of the cell total. The value of n is 2, but the value of k is highly sensitive. Researchers will be provided the values for p and k on a need-to-know basis. These values are never published and may not be shared. The calculated values (see formulas below) are referred to as *concentration ratios (e.g., "check the concentration ratios for Sample A to see if the p% rule is violated")*.

The p% and (n,k) rules are used to protect against attribute disclosure of data such as payroll, employment, etc., for the largest and most influential companies. Below are the formulas for the p% and (n,k) rules:

> Let x_1, x_2, …, x_N represent the contributions, in descending order, from respondents 1 through N with x_1 representing the largest contribution (in absolute value) and X representing the sum of the absolute values of the N individual firm contributions.
>
> The p-percent (p%) rule states that if the firms with the two largest values of a variable are removed from a cell, the cell total must be at least p% of the largest value to pass disclosure review. Let X be the aggregate total (weighted or unweighted in the same way as the requested estimates) and x_1 and x_2 be the largest two unweighted values in absolute value:
>
> $$X - |x_1| - |x_2| \geq \frac{p}{100}|x_1|$$
>
> The (n,k) rule takes the sum of the n largest contributions of a sample in absolute value. That value must be no more than k% of the cell total to be considered safe:
>
> $$|x_1| + |x_2| + \cdots + |x_n| \leq \frac{k}{100}X$$

Suppose all values of a variable are positive, n=2 and 100 – k = p, then a cell passing the (n,k) rule implies it also passes the p% rule. However the converse is not true: a cell that passes the p% rule may still fail the (n,k) rule. In practice, reporting the results for both ratios is best; the Census-vetted disclosure statistics program does this automatically. If a cell does not pass the relevant rule, output including that cell may not be released. Most magnitude values are expected to be non-negative, so any negative values in the data should be confirmed. For example, value-added and total sales could be negative, but employment should not be. If a dataset includes confirmed negative values **at the firm level**, use the absolute values to run the calculations for the p% and (n,k) rules.

For economic data from before 1992, use the (n,k) rule; for data from 1992 and later, use the p% rule to test output for possible disclosure risk. For samples spanning pre-1992 through post-1992, use both rules.

One particularly important note is that **tabular output for cells that the Census Bureau has already suppressed in existing tabular output will not be approved for release**. Checking this is the researcher's responsibility. Please speak with your RDC Admin or DAO for guidance.

V.C.1. Magnitude Concentration Ratios

In practice, the disclosure risk from over-concentration can come in several forms relevant to empirical researchers who work with economic/business data. A *magnitude variable* is a continuous variable that sums to a meaningful total across firms such that one firm could potentially dominate the total. Disclosure avoidance guidance will often refer to the concentration ratios for such variables as *magnitude concentration ratios* (MCRs). Aside from cases involving regression model output or person-level analysis (see Subsections V.C.2. Regression Model Output Derived from Economic Data Sets and V.C.3. Person-Level Analysis using Business Data), researchers need to calculate MCRs for "all magnitude variables used", which refers to any underlying/raw/untransformed magnitude variable related to a released estimate. For example, if you are requesting the mean of firm age, mean of establishment employment, and average county-level payroll per employee in industry A (i.e., payroll in industry A divided by employment in industry A), you would need to check MCRs for firm-level employment, firm-level payroll from plants in industry A, and firm-level employment from plants in industry A.

Note that "all magnitude variables used" includes external variables at the establishment or firm level. For example, say you want to release coefficient estimates for a county-level variable of manufacturing employment. If the manufacturing employment variable is publicly available at the county level, e.g., from the County Business Patterns data, concentration ratios would *not be required*. However, if you are aggregating establishment or firm level manufacturing employment from a public data source to the county level, the relevant firm level concentration ratios for firm-level manufacturing employment *would be required*.

For non-count measures based on economic data, firm counts and concentration ratios are required. If the requested output only contains count-based estimates using economic data (e.g., number of firms making investments each year, number of establishments in industry Y), then only firm counts would be required as part of the disclosure statistics for these estimates. **Concentration ratios are calculated on the original magnitude variables as provided in the dataset, not on transformed measures, e.g., you would use total value of shipments instead of *ln*(total value of shipments)**. One exception is if you recode a magnitude variable (e.g., revenue) into a binary/categorical variable (e.g., a dummy indicating whether or not a firm reported positive revenue), concentration ratios for the underlying variable would not be required unless that magnitude variable is used as a continuous variable elsewhere. If such a binary indicator is interacted with a different magnitude variable (e.g., employment), then concentration ratios for the interacted magnitude variable (employment) would be required for each respective category of the indicator variable (positive revenue or not).

In addition, concentration measures are required for the underlying magnitude variables that are used to generate any other continuous output variables, even if the underlying magnitude variables are not directly part of the output. For example, if you use x, y, and z from an internal dataset to generate Total Factor Productivity (TFP) - concentration analysis on x, y, and z should be run if estimates related to TFP are being released. **In many cases the concentration analysis for the newly created variable (TFP in the above example) would also be necessary.**

Concentration ratios are always calculated at the firm level even if the analysis is performed at the establishment level or at higher levels of aggregation (e.g., industry level). For estimates based on pooled years, that involves summing values by firm across applicable years (i.e., one record per firm for the disclosure statistics). Disclosure statistics are required for all relevant samples (explicit and implicit) and cells. If the analysis or a particular variable is at a level more highly aggregated than the firm (e.g., industry level), then the concentration ratios would still pertain to the underlying sample of firms contributing to the estimate. For example, if there is a variable for industry-level average employment and you want to report the mean of that variable using all industries in the data, the full sample firm-level MCRs for employment would be sufficient.

If you use economic data to rank groups, geographic entities, create lists of the “Top X”, etc., then full disclosure statistics are required for each reported group, and collectively for the set of excluded groups that constitute an implicit sample (e.g., the full sample minus the top X aggregations that would be reported). For example, suppose you are ranking industries by average wage and want to report the Top 5 Industries (A, B, C, D, and E). You would need to provide/check the following disclosure statistics:

- Unique firm counts and magnitude concentration ratios for industry A
- Unique firm counts and magnitude concentration ratios for industry B
- Unique firm counts and magnitude concentration ratios for industry C
- Unique firm counts and magnitude concentration ratios for industry D
- Unique firm counts and magnitude concentration ratios for industry E
- Unique firm counts and magnitude concentration ratios for all unranked industries collectively (i.e., every observation in the full sample that is not in ranked industry A, B, C, D, or E.)

Note: In general, the following data sets only require firm counts and do not, *on their own*, require concentration ratios: Annual Business Survey (ABS), Annual Survey of Entrepreneurs (ASE), Commodity Flow Survey (CFS), and Survey of Business Owners (SBO).

V.C.2. Regression Model Output Derived from Economic Data Sets

For empirical social science researchers at Census or in the FSRDCs, disclosure requests often involve regression output tables consisting of coefficient estimates, standard errors, and rounded sample sizes. For such tables, the disclosure risks are typically lower than for descriptive output or raw counts. An exemption from the "concentration analysis for all magnitude variables used" requirement is appropriate if the following conditions are met:

1. The estimates are coefficient estimates, standard errors, and/or basic model statistics (e.g., F-stat, r-squared) from multiple linear regression analysis; ***and***
2. Either (a) the table does not include every coefficient estimate from the model, or (b) the model includes at least one continuous variable as a regressor

If both conditions are satisfied and explicitly documented in the disclosure request materials, the following are sufficient for disclosure review of output derived from economic data:

- For unweighted regressions, disclosure statistics need to include unique firm counts for each relevant sample/subsample/cell as well as observations-in-sample concentration ratios (OCRs) for each relevant sample/subsample/cell
- For weighted regressions, disclosure statistics need to include unique firm counts for each relevant sample/subsample/cell as well as concentrations ratios for the weighting variable for each relevant sample/subsample/cell

The OCRs can be calculated using the usual p% and (n,k) formulas where the *key variable* (the variable for which over-concentration is being assessed, i.e., the X in the p% and (n,k) formulas) is the number of observations per firm. For example, if the regression is establishment-year analysis, the OCRs would reflect the concentrations of establishment-year observations per firm. If the regression is county-industry-year level aggregated up from establishment-year level observations, the OCRs would likewise reflect the concentration of establishment-year observations per firm. If the regression is at the firm level, with every firm in the sample contributing exactly one observation, then every firm has a value of 1 for the purpose of checking the concentration ratio, and the disclosure statistics pass trivially.

Checking OCRs ensures the observations in a cell (e.g., manuf_dummy = 0 or manuf_dummy = 1) are not dominated by one or two firms even when the count of unique firms in that cell is sufficiently large. In such a case, the regression results would be overly representative of that firm or those firms. For weighted regression, you need to confirm that the weighting variable isn't over-concentrated at a very small number of firms. For example, if the regression is weighted by employment, just check the MCRs for employment. Concentration ratios for weights are required even if the weight would not otherwise be considered a magnitude variable.

The disclosure reviewer may request additional disclosure checks. Types of output (from models or not) that do not explicitly meet the criteria discussed in this subsection would still require concentration analysis for all magnitude variables used in generating such estimates. For atypical output, the assigned DAO can be consulted regarding the potential need for additional disclosure checks. Producing OCRs, unique firm counts, and MCRs (i.e., the traditional CRs) that all meet established thresholds will likely be sufficient for the disclosure review.

V.C.3. Person-Level Analysis using Business Data

A special case arises when you use linked person-employer data to perform analysis on persons rather than firms or establishments. a hypothetic sample of 1,000 workers. Suppose 950 of these workers were employed by one firm and each of the 50 remaining workers was employed by a different firm. Suppose further that all firms have approximately 1,000 employees. A magnitude concentration ratio would not indicate any problems here; however, reported estimates for this set of workers could reveal sensitive information about the dominant firm.

If you study the workers at a sample of firms (or establishments) in relation to some type of firm (or establishment) outcome variable, then traditional concentration ratios are still required because you are producing firm statistics. However, if the sample is chosen to describe the population of workers and the outcome variable is data collected at the person-level, then the following disclosure stats should be provided instead of traditional concentration ratios:

- Individual person counts
- Worker-in-sample concentration ratios (WISCRs); i.e., concentration ratios based on the number of worker-year observations in the sample employed by each firm

Note that WISCRs also apply to any analysis using indirect worker-related observations (e.g., job-quarter observations or unemployment spell observations). These concentration ratios effectively describe the concentration of worker-related observations among the employers in the sample. The disclosure concern here is that the statistics for such samples of workers could pose a disclosure risk if the workers are highly concentrated among a small number of firms. You must check WISCRs for all samples and cells related to released estimates when business data are involved. For example, say you are requesting coefficient estimates for two indicator variables in a regression – whether the worker's establishment is a manufacturing plant, and whether the worker is a woman. You would need to provide WISCRs for the full analytical sample, splits by manufacturing = 0 and manufacturing = 1, as well as sex = 0 and sex = 1.

V.C.4. Concentration Disclosure Risk Assessment Summary

For quick reference, here are several important points regarding concentration ratios:

- ➔ Using economic data for statistical analysis typically requires providing concentration ratios for disclosure review.
- ➔ Calculate the required concentration ratios (OCRs, MCRs, and/or WISCRs) for ***all relevant samples and cells*** related to estimates derived from establishment or firm microdata.
- ➔ Calculate concentration ratios using underlying variables rather than transformed variables.
- ➔ The p% rule applies to statistics in cells that represent individual years 1992 and later or multiple (pooled) years involving the year 1992 and later.
- ➔ The (n,k) rule applies to statistics in cells that represent individual years 1991 and earlier ***OR*** multiple (pooled) years involving any year prior to 1992 (e.g., a table from data representing both 1991 and 1992 would use the (n,k) rule).
- ➔ The values of p and k are confidential. Those values are highly sensitive and may be revealed only to individuals with Special Sworn Status who have an explicit need-to-know.

The following table summarizes which disclosure statistics are typically required for an analysis:

Unit of observation	**Is firm/establishment data used?**	**Type of Output**	**Required Disclosure Statistics**
Firm/Establishment	YES	Model output meeting the criteria specified in Subsection V.C.2. Regression Model Output Derived from Economic Data Sets	- Unique firm counts - OCRs for unweighted regression; MCRs for weighted regression using the weighting variable as the key variable
Firm/Establishment	YES	Any other type of output not explicitly meeting the criteria specified in V.C.2. Regression Model Output Derived from Economic Data Sets	- Unique firm counts - MCRs for all magnitude variables used (see Subsection V.C.1. Magnitude Concentration Ratios)
Person	YES	Any	- Unique person counts - WISCRs (see Subsection V.C.3. Person-Level Analysis using Business Data)
Person	NO	Any	- Unique person counts
Indirectly Person (e.g., job level or unemployment spell level analysis)	YES	Any	- Unique person counts - WISCRs (see Subsection V.C.3. Person-Level Analysis using Business Data)

The RDC Admin and/or DAO may request different disclosure statistics in cases of heightened disclosure risk and/or scenarios falling outside the scope of the enumerated guidance. If you are concerned or unsure of the required disclosure statistics for your output request, please consult with your RDC Admin or DAO.

VI. Releases Involving Geographical Areas with Small Populations (GASPs)

Restrictions on the public release of estimates from Geographical Areas with Small Populations (GASPs) protect against the disclosure risk that accompanies record linkage from external data. GASPs have been determined to have a higher disclosure risk, and thus have become first in line to require noise injection. GASPs were called "substate" in some previous versions of disclosure guidance.

VI.A. Definition of a GASP

A GASP is defined as the union of one or more part-state geographies with a total population less than the least populous U.S. state at the time that the data were collected. "Part-state" refers to geographic entities defined at a level lower than an entire state (e.g., metropolitan statistical areas (MSAs), counties, or tracts). The definition of GASP has two key components:

G
A } Geographic selection of a part-state geography for analysis

S
P } Selected geography is less populous than the least populous contemporaneous state

An example of a potential GASP occurs when analysis is done on one city, or one county, with a small population that is contained within a state (e.g., a sample that contains only observational units in New Orleans Parish). However, the geographical area does not need to be contained within a single state to qualify. A collection of counties that span multiple states could be considered a GASP. For example, an MSA that includes parts of multiple states may be a GASP, such as the collection of counties in **Figure XI.A.1** below, whose total population is smaller than that of the least populous state.

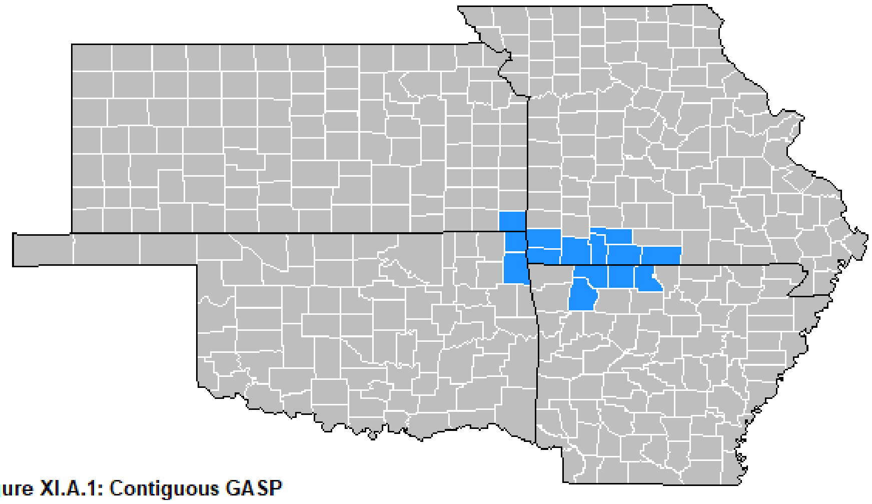


Figure XI.A.1: Contiguous GASP

Nor does the geographical area need to be contiguous to qualify as a GASP, as seen in Figure **XI.A.2**.

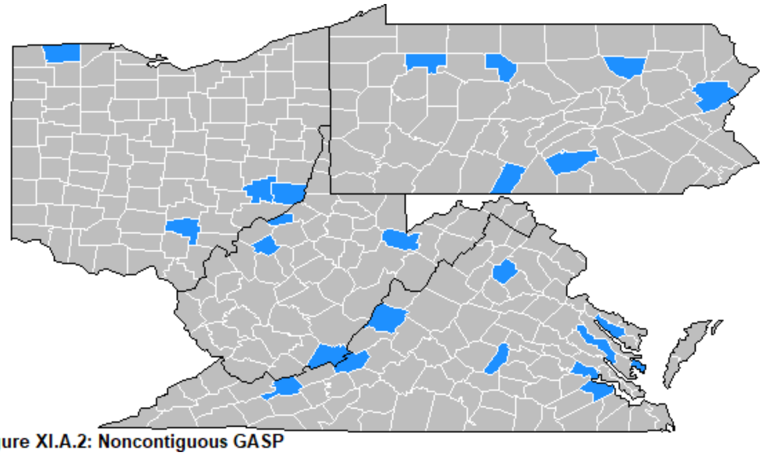


Figure XI.A.2: Noncontiguous GASP

GASP criteria also apply to implicit samples (i.e., geographic slivers that fall below the population threshold when subtracted from a parent sample are also considered GASPs). For instance, consider the following case:

Sample A: The state of Louisiana

Sample B: All parishes in Louisiana, excluding New Orleans Parish.

Implicit Sample: New Orleans Parish

The implicit sample (New Orleans Parish) would be considered a GASP if the same analysis is done on samples A and B.

Researchers interested in individual county estimates (e.g., mean for county X), should note that only about 4% of counties meet the GASP threshold on their own, so a nationwide map of county estimates could only show approximately 100 counties depending on the year and sample. The next section explains how to determine which geographic entities are GASPs.

VI.B. Population Analysis

Population analysis is a comparison of a geography's population to the population of the least populous state for the same period. To perform a population analysis, first find the population of the geographic area used in the proposed output. In cases such as a researcher-defined collection of counties or tracts, sum the population of the respective parts of the lowest level of geography for which output is being released. Compare this total population to the population of the contemporaneous least populous state.

Population analysis is required for all explicit samples defined by part-state geographies. If the estimates are national level, then population analysis is typically not required. For example, a table showing the percentage of tracts that fit into certain categories of a tract-level variable does not need population analysis if the sample is national. Similarly, a regression consisting of only continuous variables at the tract level does not require population analysis if the sample is national. If this regression had a binary indicator determined at the tract level, population analysis would be needed for each category.
Population analysis on explicit or implicit samples is required if the sample is created using publicly

known information to determine the part-state geographies in sample. For this purpose, the identities of the geographic entities are considered known if they could be derived from any data available outside the Census Bureau firewall. So, for example, if your analysis is on a collection of tracts with some characteristic and the Census Bureau publishes that characteristic for each tract, that collection is considered known and requires population analysis. If your analysis is on a collection of counties selected using an external dataset that you have brought inside the firewall, then it requires population analysis. Population analysis is required even if the dataset used for the geographic selection is not publicly available, e.g., if it was obtained through a data-sharing agreement with another government agency. If the geographic entities in sample are not known outside of the Census firewall and will not be disclosed, population analysis is likely not required. For example, if an implicit sample of geographic entities is created because individuals or establishments with missing data drop out, population analysis is likely not required so long as the names of part-state geographies that drop out are not publicly known or derivable nor made known by the researcher. If the implicit sample is public knowledge (e.g., if counties X, Y, and Z were subject to the policy in Sample A but not subject to the policy in Sample B), then population analysis would be required for explicit and implicit samples.

Additional examples when population analysis would be required:

- The sample of geographic entities is publicly known (e.g., border counties, counties subject to regulation R).
- The sample consists of a group of cities that are explicitly listed (e.g., “our sample contains observations from Cities A, B, C, and D”).

Other cases where population analysis is likely not required include:

- The sample of geographic entities is based on having at least one plant of size X where X is only known from the internal data.
- The analysis is nationally representative and does not have a set list of part-state geographic entities from which the sample is derived.

Researchers using internal data to define samples, categories, etc. based on geographic entities should provide justification that the samples or categories cannot be ascertained from publicly available information. If the DAO assesses that the geographic entities in-sample could be approximated using public data, they may request population analysis even if the definitions are made based on internal data only.

Geographically defined regression categories require additional population analysis if such estimates are being released. This includes:

- Geographically defined categorical variables included by themselves or crossed with continuous variables.
- An interaction term involving more than one categorical variable, one of which is geographically defined – analysis is needed on the cross tab of all geographic categorical variables.
- A dependent variable that is geographic and categorical – analysis is needed between it and any geographically defined independent categorical variable.

Examples of regression variables that likely require additional population analysis are:

- In county X or not
- In a metro area or not in a metro area
- In a county with regulation X or not in one
- In a county that is > x% of a certain race
- Urban or rural
- Housing unit within X miles of a particular city center or not
- Establishments in counties within X miles of a particular county

Examples of regression variables that do not require additional population analysis are:

- average income defined at the tract level
- percent race X at the tract level

Again, population analysis is required only if the information used to generate the categories is publicly available. For example, consider a case where the regression analysis is being performed on a sample of households from four cities in New York – Albany, Buffalo, Rochester, and Syracuse – with some categorical variables defined at the block level (e.g., does the block have a median income above Y). Suppose these block-level variables are based *solely* on internal data. Population analysis would then be required only for the collection of cities.

If the data span multiple years, the population analysis is done by taking the sum of the individual year populations of that region. Similarly, the threshold is the population of the least populous state summed over the corresponding years.

The population analysis will either pass (i.e., exceed the smallest state population threshold) or fail. Products that pass population analysis can be released directly by a DAO, barring any other complicating features. Products that do not pass population analysis will need to go before the DRB and will likely require noise injection or another approved methodology. Sign & significance output is subject to the same rules regarding population analysis as numeric output (see Section III.D. Sign and Significance).

To establish that a geographic area meets the population threshold and is not a GASP, you need not show the exact population of the area. For example, if your geographic area is the union of four MSAs including the Toledo MSA and no output will be released on the individual MSAs, then it suffices to note, “The sample includes the Toledo MSA, whose population alone is XXX,XXX, which exceeds the GASP threshold.” Obviously, the total population of the four MSAs exceeds this, so you have established that your sample is not a GASP.

VI.C. Model-Based Estimates

A *model-based estimate* is defined here as an estimate derived from an analytic process such as an analysis of variance, a fixed-effects regression, or a factor analysis. Some model-based estimates can be released without noise injection. Examples include:

- Model estimates where all Census Bureau data used in the model appear on the **Exempt Sub-State Geographies Information Products List**
- Model estimates that contain a **pooled component** that is based on areas as populous as the smallest state, and in which the *weight on the direct component goes to zero as the number of*

sampled entities in the cell goes to zero. Small-area estimation models often fall into this category.

- Model-based estimates that have DRB recommendation for approval, and have been approved by the Chief Scientist and DSEP

To confirm that a model-based estimate does not require noise injection, you will have to submit a detailed description of the model, which will need to be approved by the DRB.

Below are a few examples of models that potentially fit the conditions (when including parameters defined for geographic areas that are at least as populous as the smallest state):

- Random effect models
- Multilevel Regression and Post-stratification (MRP) models
- Bayesian or maximum likelihood hierarchical models
  - ➔ Note: For hierarchical models, only the estimates at a sufficiently populous level and at the level immediately preceding it may be released

VI.D. Noise Injection

If output based on a GASP does not meet the above conditions for direct and model-based estimates, the DRB will release estimates for sub-national geographies only with an appropriate noise injection method. This applies to all estimates, both direct and model-based estimates, including counts in tables, continuous variables in microdata, and economic magnitude data (which means that noise must be injected in addition to legacy cell suppression techniques). The injected noise can be formally private or non-formally private. Please see the guidance on differential privacy and noise injection for more information.

The Center for Enterprise Dissemination-Disclosure Avoidance (CED-DA) will work with you on the noise injection process. A consultation with CED-DA should be scheduled once it is determined that your output will require noise injection. All noise injection releases must then go through the DRB review process. In many cases, we will need an external expert review of novel noise injection techniques, as well as the approval of the Chief Scientist. This could require considerable time and resources, so please identify the need to use noise injection early in the research process.

VII. Reporting Match Rates

VII.A. Match Rates

If datasets are matched to each other and the match rate is reported, the rate must be top-coded at 99.5%, i.e., you may report that the match rate is 99.5% or higher without giving any more detail. The rate may be reported only if at least 10 records matched and at least 10 records did not match. Match rates must be rounded according to the unweighted proportion rounding rules described earlier.

If the requirement for at least 10 non-matches is not met, the researcher can state that the match rate is "greater than X%," where X% of the sample and (100-X)% of the sample each represent at least 10 records. Similarly, if the requirement for at least 10 matches is not met, then the researcher may say the match rate is "less than X%," where X% of the sample and (100-X)% of the sample each represent at least 10 records.

VII.B. Protected Identification Key (PIK) Rates

PIK rates do not need to be top-coded but must be rounded according to the unweighted proportion rounding rules described earlier.

VIII. When Estimates Do Not Meet Criteria for Release

For all disclosure requests, it is critical to document that all relevant disclosure requirements are met. This includes clear supporting documentation as well as providing all required disclosure statistics. Some of these criteria were discussed in earlier sections of this handbook while others will be subsequently described in the proceeding sections.

If the RDC Admin or DAO finds problems or estimates that do not meet criteria for release, the researcher will be asked to do one or more of the following things:

Collapse—that is, combine—certain cells. This will avoid disclosure problems at the expense of output detail and is the preferred course of action.

Suppress the numbers in the affected cells. All suppressed cells should be marked with a "D" with a note explaining those cells were suppressed for disclosure reasons. You will almost always need to carry out complementary (secondary) suppression on other cells to prevent a data user from being able to manipulate the released cells to determine the values of cells that needed to be suppressed. Complementary suppression is by far the more difficult and time-consuming part of disclosure analysis. The DAO will only allow release of relatively simple tables (such as those generally found in journal articles) for which complementary disclosure can be carried out simply. High frequency of suppressed cells and/or complicated complementary suppression considerations will require DRB review.

Reconsider the output, by asking what you are trying to show and alternative ways to convey that information besides tables. For example, you may be able to summarize the information in the cells rather than showing all the cells.

IX. Special Considerations for Particular Surveys

IX.A. American Community Survey

The ACS is not designed to produce block-level estimates. The DRB, FSRDC Disclosure Avoidance Officers, and the ACS reviewer for FSRDC proposals discussed the appropriate use of Census block-level ACS data in FSRDC projects. They established the following guidelines.

Descriptive statistics/tables

- Block-level tabulations are not permitted.
- Tabulations of researcher-defined areas using Census block IDs will be carefully reviewed, especially for:
  - ➔ Relationship to any other geographic variables used in the project
  - ➔ Relationship to other published data products
- Such tabulations may be subject to DRB review.

Model-based statistics
Typically, block-level information is allowed in models. However, it may be barred for particular projects if specific disclosure concerns arise during proposal or output reviews.

The following can be allowed:

- **Block-level contextual variables in an individual-level or household-level model**

This includes variables such as:

- Percent white in person's block.
- An indicator variable for whether a household's block is within a certain area – possibly subject to the GASP policy discussed in Section VI. Releases Involving Geographical Areas with Small Populations (GASPs).

Researchers who use block-level contextual variables will have to justify why they cannot use higher levels of geography like block group or tract. Using blocks in this way will typically not present a disclosure problem if the analytical sample consists of households from the entire nation. However, samples for lower levels of geography may introduce concerns and would be subject to the GASP policy (see Section VI. Releases Involving Geographical Areas with Small Populations (GASPs)).

- Summary statistics and tables describing the analytical sample will be carefully considered, as described above.
- As customary for disclosure review of model output, cell counts for any dummy variables must meet the relevant threshold (usually 3+ unique persons; see Section V.A. Cell Size Thresholds). Furthermore, the model should contain at least one continuous independent variable.
- The finest level of detail that may be shown for Group Quarters data is Institutional/Non-institutional. There are no exceptions to this rule, which applies to all years.

IX.B. Longitudinal Employer-Household Dynamics

For the Longitudinal Employer-Household Dynamics (LEHD) data available in the FSRDCs, researchers should use a mixture of the rules for economic and demographic data. Specifically, the standard disclosure rules apply for either person-level or business-level analysis. This means you should use the threshold (count) or concentration (p% or (n,k)) rules described above. See Section V.C. Disclosure Risk from Over-Concentration for more information. Here are some scenarios and the disclosure statistics required in each:

- For person-level analysis using no firm/establishment data, only counts are required
- For person-level analysis that also use firm/establishment data, person counts, and the workers-in-sample concentration ratios are required (see Subsection V.C.3. Person-Level Analysis using Business Data).
- For firm- or establishment-level analysis using T13 components of LEHD only, firm disclosure statistics should be calculated using SEIN.
- For firm- or establishment-level analysis using T26 components of LEHD, firm disclosure statistics should be calculated based on the entity level used for the analysis. If SEIN or establishment is used for employer characteristics, then use SEIN for the disclosure stats. If the alpha firm id is used, then use that for the disclosure stats.[5]

Per the data use agreements with states, all results to be disclosed must include data from **at least three states**, unless your project has obtained a specific exemption to this rule during proposal review. You need to include the number of states for each sample, subsample, and implicit sample in your disclosure statistics. Models may include geographic controls for more detailed geographic levels, but the coefficients on these controls may not be reported. It is okay to note on the table of coefficients: “includes controls for [insert geography]”.

Under the agreements that allow the Census Bureau to use their data in the LEHD program, some states must review research output before the output is released publicly. Please contact your RDC Admin or DAO about these requirements and plan for any needed extra review time.

IX.C. UMETRICS

For research using UMETRICS data, all tables, figures, and summaries in the requested disclosure material must contain information from at least three universities. Model results, coefficients and standard errors of university-specific controls may not be released (this is similar to the LEHD three-state rule).

IX.D. Medical Expenditures Panel Survey – Insurance Component

For the MEPS-IC, if research is conducted using only the public sector, that output is not subject to Title 13 and thus no disclosure review is required. However, if the requested estimates involve any private sector data from the MEPS-IC or any other title-protected data, the usual disclosure stats are required. When using both the public and private portions of MEPS-IC, all estimates related to the private sector

[5] Note that when Title 26 portions of the LEHD are used, the source of the information should be taken into consideration when possible in considering the minimum number of records. For example, a single IRS Form 1040 can pertain to multiple people, while a W2 would only pertain to one person.

must pass disclosure review. Where totals across the public and private sector are reported, you may have to suppress some statistics related to the public sector to prevent the implicit release of information related to the private sector. Consider the following example.

Suppose this table is requested for release:

Citizen (private + public)	Private payroll	Public payroll
100,000	45,000	55,000

If Private fails concentration ratios, then the Citizen (or total) would need to be suppressed as well.

Citizen (private + public)	Private payroll	Public payroll
D	D	55,000

IX.E. Social Security Administration Data Sets

Some data available to researchers come from the Social Security Administration (SSA), either directly or indirectly. After a data product using an SSA dataset receives DRB approval (whether by the DRB itself or by a DAO) and a DRB clearance number, the output is not ready for release until SSA approves.

First, the researcher must prepare a form, the SSA Output Information Form (SSA-OIF), consisting mostly of a brief description of the research output. To minimize delays later, the researcher is encouraged to prepare the SSA-OIF before the disclosure review is started. The DAO will route the form and the DRB-cleared output to SSA, which has 24 to 72 hours to raise any disclosure-related objections. If SSA does not have any objections, then the output is fully cleared for release and may be shared like any other cleared data product, always displaying the DRB clearance number.

In filling out the SSA-OIF, researchers should explicitly note whether the analysis is weighted or unweighted, as SSA often asks for that information if it is not provided.

The datasets requiring an OIF are as follows[6]

- Numident
- SSA Master Beneficiary Record (MBR)
- SSA Supplemental Security Record (SSR)
- SSA Payment History Update System (PHUS) datasets
- Household Composition Key (since it is created using the Numident)
- LEHD Individual Characteristics File (ICF) (since it is created using the Numident)
- Household Survey Link

[6] Of these datasets, the Household Survey Link requires a 72-hour SSA review period, while the others require a 24-hour SSA review period.

X. Modern Methods

Most disclosure methods used until now aim to make it more difficult for a data intruder to reveal sensitive information about a respondent. However, these methods are no longer sufficient in protecting Census data from attacks by intruders with sufficient knowledge, computing power and auxiliary information. Traditional disclosure methods rely on intruders not having these resources.

All output released from any data provider—such as the Census Bureau—gives away some privacy about the underlying dataset. A new class of methods, known as *formally private* methods, allow the data provider to quantify a mathematical guarantee of how much privacy could be lost and limit that privacy loss as desired.

Differential privacy is the most common type of formal privacy protection. Differential privacy is not itself a method, but rather is a criterion that a method must satisfy in order to be acceptable. Furthermore, saying that a method is differentially private is a statement about the algorithm used to protect the data rather than about the final data released.

More information on modern methods will be added in future versions of this Handbook. For now, here is a list of references for those interested in learning more:

- A non-technical introduction to the major ideas in the differential/formal privacy literature: https://dash.harvard.edu/handle/1/38323292

- John Abowd, Ian Schmutte, William Sexton, and Lars Vilhuber developed a list of readings to help economists learn about differential privacy:
https://labordynamicsinstitute.github.io/privacy-bibliography/

- The Dwork-Roth monograph is the nearest thing to a textbook on differential privacy. Chapter 1 provides a gentle, prose introduction to the topic, and Chapters 2 and 3 cover most of the foundational theorems common to the literature:
https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf

- Jinshuo Dong wrote a very nice blog post discussing the hypothesis-testing approach to semantic interpretation of the formal privacy guarantee: https://dongjs.github.io/2020/01/15/Privacy.html

- Those who find the approach to semantically/concretely interpreting the privacy guarantee appealing might also appreciate the growing number of papers that adopt a similar perspective: https://arxiv.org/abs/0811.2501 (the seminal paper) and https://arxiv.org/abs/1905.02383

References

J. Dong, "How Private Are Private Algorithms?" Jinshuo's Blog, <https://dongjs.github.io/2020/01/15/Privacy.html>

J. Dong, A. Roth, W. Su, "Gaussian Differential Privacy," arXiv preprint arXiv:1905.02383, 2019.

C. Dwork and A. Roth (2014), "The Algorithmic Foundations of Differential Privacy", Foundations and Trends in Theoretical Computer Science, Volume 9, No. 3–4, pp. 211-407. <https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf>.

Federal Committee on Statistical Methodology, "Report on Statistical Disclosure Limitation Methodology, Second version," Statistical Policy Working Paper 22, U.S. Office of Management and Budget, Washington, DC, 2005. <https://nces.ed.gov/FCSM/pdf/SPWP22_rev.pdf>.

L. McKenna and M. Haubach, "Legacy Techniques and Current Research in Disclosure Avoidance at the U.S. Census Bureau," Research and Methodology Directorate, April 2019. <https://www.census.gov/library/working-papers/2019/adrm/legacy-da-techniques.html>.

A. Reznek and T. Riggs, "Disclosure Risks in Releasing Output Based on Regression Residuals," American Statistical Association, Proceedings of the Section on Government Statistics and Section on Social Statistics, 2005, pp. 1397-1404.

L. Wasserman and S. Zhou, "A statistical framework for differential privacy," Journal of the American Statistical Association, Volume 105, No. 489, 2010, pp. 375-389.

A. Wood, et al, "Differential Privacy: A Primer for a Non-Technical Audience," Vanderbilt Journal of Entertainment & Technology Law, Volume 1, No. 1, 2018, pp. 209-276.

Appendix A: Implicit Samples

Often researchers need to release results based on a main analytical sample and one or more subsamples. Sample sizes and disclosure statistics must be provided for each sample and subsample used to create estimates. In addition, "implicit samples" are often created when subsamples are used that represent a subset of another sample or combination of samples. Implicit samples are the difference between the larger sample and its subset. These samples must also be addressed in disclosure review.

Example 1

- Sample 1: all firms in sector X throughout the US
- Sample 2: all firms in sector X in all states except California

The implicit sample created is all firms in sector X in California.

Example 2

- Sample 1: people of any age
- Sample 2: people between the ages of 0-65

The implicit sample created is all people aged 66 or over.

Figure 1 gives a visualization of how implicit samples are created in situations like those given in Examples 1 and 2.

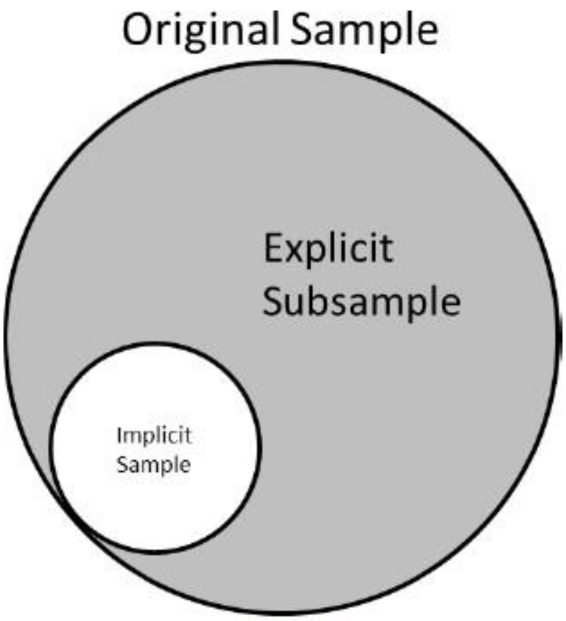


Figure 1

Correctly identifying all implicit samples can be tricky sometimes, especially if the main sample is divided in multiple ways. Researchers must work with their RDC Admin to ensure they have accounted for all implicit samples appropriately.

Example 3

- Sample 1: All Firms (n=100)
- Sample 2: Employers (n=48)
- Sample 3: Large Firms (n=30)

A couple of implicit samples are created here.

- Implicit Sample 1: Sample 1 - Sample 2 = Non-employers
- Implicit Sample 2: Sample 1 - Sample 3 = Small Firms

We can think of this in terms of a frequency table. Sample 1 is the grand total and Sample 2 and Sample 3 are marginal cells. The other marginal cells represent the two implicit samples and are in italics.

	Employers	Non-employers	Total
Large Firms			**30**
Small Firms			*70*
Total	**48**	*52*	**100**

Figures 2 and 3 depict how the implicit samples can be derived from knowing the population total along with the number of non-employers and small firms.

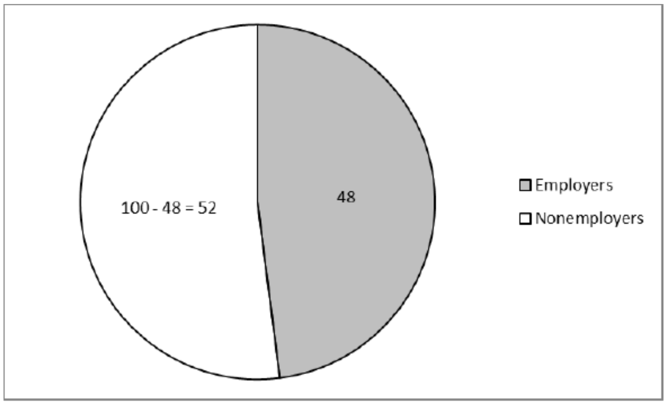


Figure 2

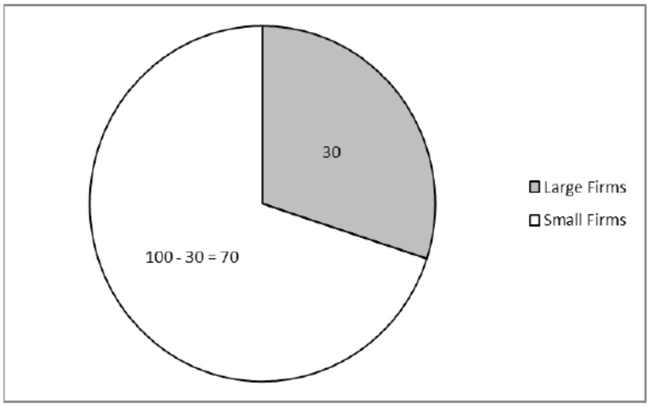


Figure 3

No implicit sample is created by comparing Sample 2 and Sample 3. The samples are *overlapping* and it's impossible to identify the size of any subsets without more information. A researcher's memo should clearly document that no implicit samples exist and the counts between the samples are not needed.

Figures 4 and 5 show two possible sets of values for the interior cells that would result in the same marginal totals. There are many other possible sets of interior cell values that would result in the same marginal totals. However, without more information, we are not able to determine the true values.

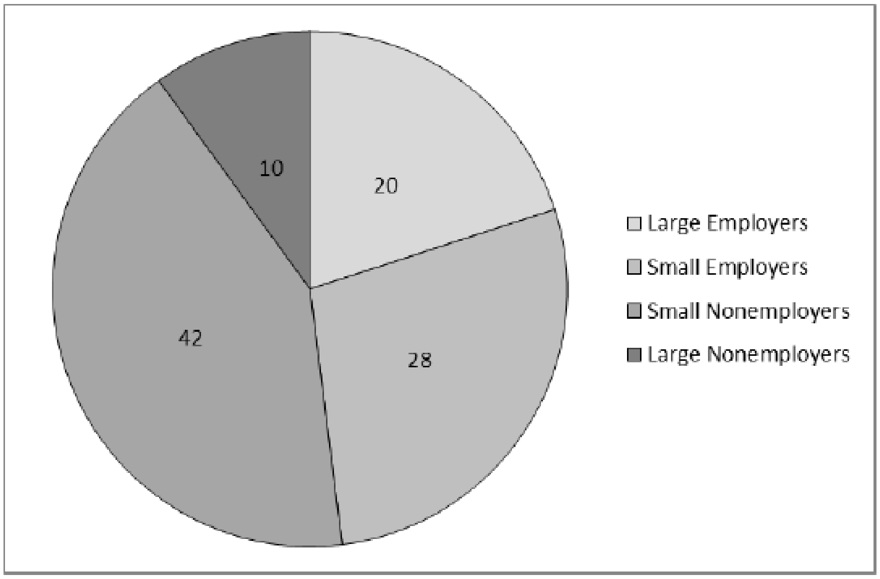


Figure 4

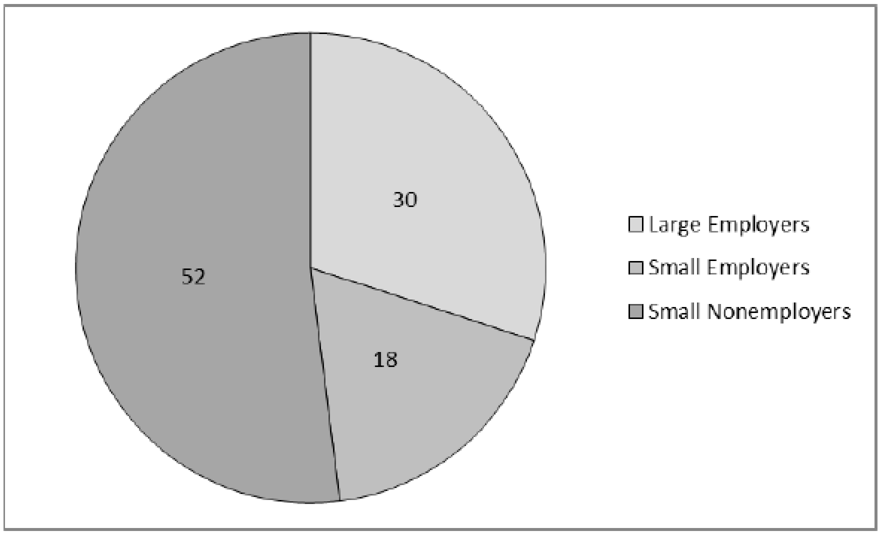


Figure 5

Example 4
One exception is a case where one of the samples has no members.

- Sample 1: All Firms (n=100)
- Sample 2: Employers (n=0)
- Sample 3: Large Firms (n=30)

This allows us to identify completely the size of all the marginal and interior cells.

Since we now know there are no employers in our sample, all large and small firms in our sample must be non-employers. Therefore:

- Implicit Sample 1: Large Employers = 0
- Implicit Sample 2: Small Employers = 0
- Implicit Sample 3: Sample 1 - Sample 2 = Non-employers
- Implicit Sample 4: Sample 1 - Sample 3 = Small Firms
- Implicit Sample 5: Sample 3 – Sample 1 = Large Non-employers
- Implicit Sample 6: Implicit Sample 4 – Implicit Sample 2 = Small Non-employers

	Employers	Non-employers	Total
Large Firms	0	30	**30**
Small Firms	0	70	70
Total	**0**	100	**100**

Example 5
Now, let’s say a fourth sample is added to our samples from **Example 3**, as such:

Sample 4: Large Employers (n=27)

One interior cell is now explicitly defined, and the other interior cells may be identified through simple subtraction from the marginal cells. Thus, there are now three new implicit samples to be accounted for in this case.

	Employers	Non-employers	Total
Large Firms	**27**	3	**30**
Small Firms	21	49	70
Total	**48**	52	**100**

To sum up, here's a list of all the samples the researcher now needs to identify.

- Sample 1: All firms (n=100)
- Sample 2: Employers (n=48)
- Sample 3: Large Firms (n=30)
- Sample 4: Large Employers (n=27)

The implicit samples are

- Implicit Sample 1: Sample 1 - Sample 2 = Non-employers
- Implicit Sample 2: Sample 1 - Sample 3 = Small Firms
- Implicit Sample 3: Sample 2 - Sample 4 = Small Employers
- Implicit Sample 4: Sample 3 - Sample 4 = Large Non-employers
- Implicit Sample 5: Sample 1 - Sample 2 - Implicit Sample 4 = Small Non-employers

Figure 6 depicts this numeric example.

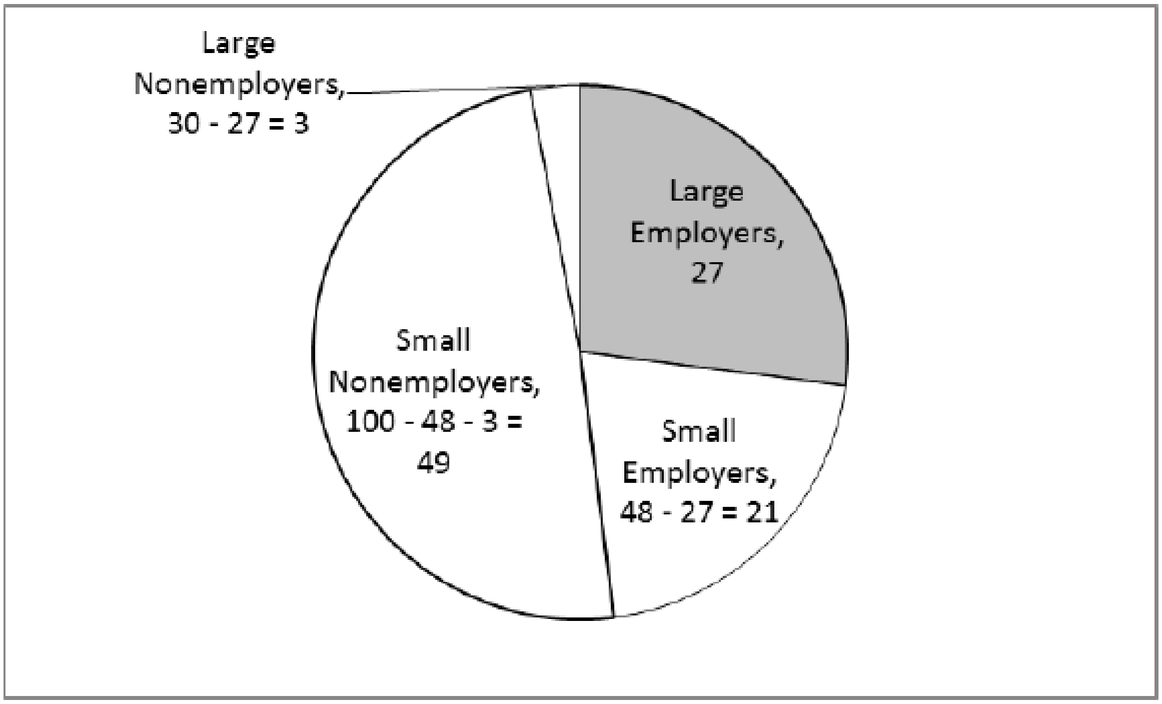


Figure 6

This example illustrates that dividing the sample in multiple ways can quickly increase the amount of created implicit samples. Researchers should keep this mind when determining samples for which they want to release estimates.

Reviewing Implicit Samples

Researchers must identify all implicit samples, including those created from different sample definitions:

- within a certain release request
- between a current request and a prior release
- between the project and other published data, either from standard publications or other RDC projects

Researchers must provide disclosure statistics for all created implicit samples (use unweighted counts). As usual, these statistics include:

- Sample sizes
- Cell counts for categorical variables
- Concentration ratios (for non-noisy data from economic datasets)

The usual disclosure criteria will apply to implicit samples. Disclosure statistics are not required when the same analysis is not being run on the samples contributing to a given implicit sample. For example, say that Sample B is a subset of Sample A. If regression specification 1 is run using Sample A and Sample B, then disclosure statistics are required for Sample A, Sample B, and the implicit sample between A and B. If regression specification 1 is run only on Sample A and regression specification 2 is run only using Sample B, then the implicit sample disclosure statistics are not required.

Appendix B: Volume of Output Examples

The examples below all stem from previously released results. All numbers and some variable names have been changed so as to not "publish" a researcher's results.

Appendix Table B1: Descriptive Statistics

Variable	mean_all	lsoft_qui_1	lsoft_qui_2	lsoft_qui_3	lsoft_qui_4	lsoft_qui_5
Exit	0.0546	0.078	0.0572	0.0494	0.0442	0.0416
Extojob	0.0182	0.026	0.0208	0.0156	0.0156	0.0156
edd1	0.0364	0.052	0.0416	0.039	0.0312	0.0234
edd2	0.0676	0.0728	0.0702	0.0702	0.065	0.0572
edd3	0.0806	0.078	0.0806	0.0832	0.0832	0.078
edd4	0.0754	0.0572	0.0676	0.0702	0.0806	0.1014
Woman	0.1118	0.1118	0.1092	0.1092	0.1248	0.1066
age_1	0.0234	0.0364	0.026	0.0234	0.0182	0.0156
age_2	0.0338	0.0364	0.0338	0.0312	0.0312	0.0286
age_3	0.0338	0.0338	0.0338	0.0312	0.0338	0.0338
age_4	0.0312	0.0286	0.0286	0.0286	0.0312	0.0338
age_5	0.0312	0.0286	0.0312	0.0312	0.0312	0.0364
age_6	0.0312	0.0286	0.0286	0.0312	0.0312	0.0338
age_7	0.0286	0.026	0.0286	0.0312	0.0312	0.0312
age_8	0.0234	0.0208	0.0234	0.026	0.026	0.0234
age_9	0.0156	0.013	0.0156	0.0156	0.0156	0.013
age_10	0.0078	0.0104	0.0104	0.0104	0.0078	0.0078
N (2012)	23500000					
N (all years)	71500000					

Volume of output: Each cell in this table (mean, quintile, and number of observations for 2012 and all years) counts towards the volume of output. There is a total of 104 cells.

Appendix Table B2: Regression Output

	dependent var1	dependent var1	dependent var1	dependent var2
	b/se	b/se	b/se	b/se
ind. Var1	0.0587***	-0.0095*	-0.0517***	-0.0014
	(0.0064)	(0.0015)	(0.0032)	(0.0025)
ind. Var2	0.0231***	0.0274***	-0.0357***	0.0126***
	(0.0078)	(0.0019)	(0.0159)	(0.0027)
ind.var3	0.9876***	0.0578***	-0.0004***	0.0188***
	(0.00541)	(0.0014)	(0.0019)	(0.0017)
ind.var1#ind.var2	0.1234***	0.0812***	0.0652	0.0224***
	(0.0032)	(0.0025)	(0.0017)	(0.0014)
r2_a	0.388	0.377	0.852	0.324
N	1610000	1610000	1610000	1610000

Each coefficient-standard error combination in this table counts as one estimate. Each adjusted R2 counts as one estimate. The number of observations is repeated across the four regressions, so it counts as only one estimate. The total number of estimates in this table is 21.

Appendix Table B3a: Other

	NAICS digit level				
	2-digit	3-digit	4-digit	5-digit	6-digit
(mean)	-543.1	-754.8	-42.15	-89.21	-0.789
(std dev)	4200	1542	1752	654.3	987.6

Appendix Table B3b: Other

Table 2. Serial correlation in measurement error					
	NAICS digit level				
	2-digit	3-digit	4-digit	5-digit	6-digit
1-year	0.012	0.012	0.012	0.012	0.258
5-year	0.045	0.045	0.045	0.045	0.058
10-year	0.654	0.321	0.456	0.987	0.852

Note, Tables B3a and B3b together count as 20 estimates. The row of standard deviations in Table B3a are in relation to the row of means and therefore do not count as independent cells.

Table B4a

Sample:	**1**	**A**	**B**	**C**	**D**	**E**
Period:	Full	1973-1977	1978-1982	1983-1987	1988-1992	1993-1997
Dep. Var.:	Log (1+Capex/CPI)					
log(labor productivity)	0.222	0.223	0.278	0.256	0.200	0.346
	0.005	**0.026**	**0.035**	**0.015**	**0.026**	**0.037**
log(plants per segment)	-0.001	NR	NR	NR	NR	NR
	0.003	NR	NR	NR	NR	NR
log(plants per firm)	-0.006	NR	NR	NR	NR	NR
	0.009	NR	NR	NR	NR	NR
plant age (/100)	0.753	NR	NR	NR	NR	NR
	0.095	NR	NR	NR	NR	NR
Industry-year fixed effects	Y	Y	Y	Y	Y	Y
Observations	2589000	165000	447000	221000	229000	278000
R2	0.0606	0.0888	0.0447	0.0221	0.0452	0.0332

Table B4a includes 21 estimates (cells). Note that the ‘Y’ (yes) indicator for fixed effects and the ‘NR’ (not reported) cells do not count.

Table B4b

Sample:	**1**	**A**	**B**	**C**	**D**	**E**
Period:	Full	1973-1977	1978-1982	1983-1987	1988-1992	1993-1997
Dep. Var.:	Log (1+Capex/CPI)					
log(labor productivity)	0.852	0.447	0.951	0.112	0.337	0.526
	0.012	**0.074**	**0.066**	**0.079**	**0.015**	**0.056**
log(labor productivity) x union	-0.245	-0.088	-0.074	0.359	-0.126	-0.749
	0.062	**0.111**	**0.333**	**0.444**	**0.556**	**0.667**
Union	0.155	0.397	0.179	0.247	0.397	-0.128
	0.553	**0.668**	**0.331**	**0.447**	**0.115**	**0.226**
log(plants per segment)	NR	NR	NR	NR	NR	NR
	NR	NR	NR	NR	NR	NR
log(plants per firm)	NR	NR	NR	NR	NR	NR
	NR	NR	NR	NR	NR	NR
plant age (/100)	NR	NR	NR	NR	NR	NR
	NR	NR	NR	NR	NR	NR
Industry-year fixed effects	Y	Y	Y	Y	Y	Y
Observations	2589000	165000	447000	221000	229000	278000
R2	0.0405	0.0123	0.0860	0.0158	0.0748	0.2500

Table B4b includes 24 estimates (cells) that count towards volume of output. Note that the number of observations across Table B4a and B4b are the same so they only count once.