Federal Statistical Research Data Center Disclosure Avoidance Methods:

A Handbook for Researchers

VERSION 2.1.0

The disclosure avoidance methodology in this document comes from guidance by the U.S. Census Bureau's Disclosure Review Board. This document was released on August 12, 2021. Changes to this guidance will occur. Please consult your RDC Administrator 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 an FSRDC researcher with Special Sworn Status (SSS), you have taken an oath to protect sensitive and confidential data. Any products created from internal data are confidential until they have undergone disclosure review and have been approved for release. The Data Stewardship Executive Policy Committee (DSEP) oversees data stewardship activities and sets rules and policies that all Census staff and all people 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 (DAR) and Disclosure Avoidance Officers (DAO) 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 Federal Statistical Research Data Center (FSRDC). Output includes, but is not limited to, statistical output, notes, programs, graphs, and tables or statements on sign and significance of estimates. The process requires researchers to prepare a disclosure request. FSRDC Administrators (henceforth referred to as the RDC Administrator) have been trained in disclosure avoidance methods and work with the researcher to prepare a request for disclosure review. Once the RDC Administrator determines the requested output is ready to be reviewed by a DAO they 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 or bring requested output before the DRB. See the Disclosure Avoidance Review Procedures Handbook for more information on the disclosure review process.

This document explains disclosure avoidance review methods for commonly requested output from the FSRDCs. As the researcher, it is your responsibility to provide disclosure statistics showing that your output follows these methods. The RDC Administrator and 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.

There are two main types of disclosure avoidance methods: legacy methods and modern methods. Legacy methods include threshold or count rules, rounding, 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.

II. 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.

Note that when working with variables from Title 26 data (e.g., Federal Tax Information (FTI) and/or Internal Revenue Service (IRS) data), the internal variable names and labels are, per guidance from the IRS, "administratively sensitive". As such, the exact variable name and label cannot be released. For example, if the administratively sensitive variable name for an FTI variable is "breakfast" in the internal data, the raw variable name listed in any item being released should be renamed (e.g., "breakfast_variable"). Researchers should make such adjustments and alert their RDC Administrator that such changes have been made before the disclosure review commences.

II.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). **It is FSRDC policy 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. We have 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 down via various categories (e.g., sex=0 for male and sex=1 for female), provide the frequency counts for each category so the RDC Administrator 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 (dependent on which dataset is used). In some cases, concentration ratios must also be calculated for each category of categorical variables used in summary statistics. See Sections IV.A. Cell Size Thresholds and IV.C. Concentration Rules for more details.

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

3. We typically do not publish true percentiles (including medians), as these could potentially correspond to actual confidential values.
    - Instead, researchers are to calculate a pseudo-percentile. That is, take the mean value from the subset of observations around the percentiles—at least five observations on either side, for a total of at least 11 observations for a given quantile.
    - If you compute multiple quantiles on 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 observations 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. A special case to consider is when the level of analysis differs from the level the quantile described. For example, if the analysis is employee-level but the quantile pertains to firm production, the reported quantile would need to be derived from 11 firms rather than 11 employees.

- Any reported quantile needs to be rounded according to the Rounding Rules.

Be advised that your RDC Administrator or DAO may determine that reporting a quantile (even when calculating a pseudo-percentile) is a disclosure risk. Such a determination could 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 values for the statistic to be reported (either the minimum or the maximum).

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

II.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 could be a difficult task as researchers may be not be able to easily identify if these tabulations already exist.

While a reference file of published tables is in development, researchers should use https://data.census.gov/cedsci/ or other known resources, to identify if 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 review and 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, do not pose a threat to the privacy protections on the official product. In this case, the research output along with this

explanation and support documentation must be brought to the DRB for review.

II.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 plants belong to that specific firm(s).

All the disclosure statistics calculations (counts and concentration ratios) listed below should be done at the level of individuals, households, or firms. One set of requested output might focus on different units of analysis such as counties, families, states, etc. Though this is important for analytic reasons, 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 IV.A. Cell Size Thresholds and IV.C. Concentration Rules 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**. (Hint: 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 III. 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 pass concentration ratio rules.
- If the model has a continuous variable on the left-hand side and if 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.
- If the model has a categorical variable on the left-hand side, then make the standard disclosure calculations for all of the 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 of the firms, households, or individuals in each category, crossed by the categories on the left-hand side of the model.
- If you have interaction terms in your model constructed from binary variables, the categories for which counts and concentration ratios are required varies depending on what estimates are being released. Below are several examples of how to apply this guidance.

- ➔ Consider a regression model with binary1, binary2, and interaction12 among the independent variables. If reporting coefficient estimates for all three of these variables, disclosure stats 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 would **not** need to additionally provide disclosure stats 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, disclosure stats should be provided for the following categories: binary1=0; binary1=1; interaction12=0; interaction12=1.
- ➔ If only reporting the coefficient estimate for interaction12, the disclosure statistics should be provided 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, the disclosure stats should be provided for binary1=0 and binary1=1 regardless of whether the coefficient estimate for binary1 is being released.
- ➔ Note, using a binary outcome variable in the regression model would require the appropriate cross tabs as 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, disclosure stats should be provided 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.).

- You only need to provide disclosure statistics for binary variables that you report. For example, if you use dummies as "controls" or "fixed effects" for firms, households, individuals, or states, but do not report them, you do not need to provide disclosure statistics for them.
- 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), the disclosure statistics must be provided 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 of the counts and concentration ratios must be run on the implicit sample as well. See Section III.B. Implicit Samples for more details.
- All coefficients must be rounded to four significant digits. See Section IV.B. Rounding Rules for more details.

II.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 a successful re-identification or database reconstruction attack

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 are going to be released, the data must pass all disclosure review rules. Once any underlying data or estimates pass disclosure review and are released, you can create any graphs or figures using that released data. The guidance in this section is for graphical output where underlying data or estimates are not being requested for release. Note that whether or not the underlying data is also requested for release, you need to provide the appropriate disclosure statistics for all graphs and figures. Further, the underlying data or estimates in the graph should be provided for disclosure review in support files. For example, if the request includes graphical output depicting regression coefficients, the regression results should be included among the support files. Researchers can certainly request to release the underlying data for figures as a table instead. Such estimates would need to follow all disclosure rules (e.g., rounding). Using the approved and released underlying numbers, the researcher could then make as many figures as they want based on the approved numbers.

The preferred method for graphical output is to round the underlying data and produce the graph or figure with the rounded data or estimates. You should provide the associated underlying data, rounded values, and disclosure statistics in support files. If the underlying data passes all disclosure rules (including rounding), the resolution of the figures is inconsequential. If the underlying data or estimates cannot be shown directly in support files then 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. Further, 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 .3 and .4 provides only one significant digit of precision. The graph’s DPI needs to have a low enough resolution that no more than four significant digits are revealed. In short, if the resolution meets or undercuts these standards, the underlying data presented in the graphs or figures are not subject to rounding rules.

Graphical output counts towards volume of output (see Section VIII. Volume of Output). Reducing precision in ways described above is looked upon favorably when determining if the volume of output in a request will be approved.

All figures need to be produced in one of the following approved formats:

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

II.C.1. Kernel Density Plots

You may wish to produce supporting figures or graphs to describe the data used in your projects. One of the most common ways 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, we recommend that researchers 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. Entity counts should be reported for each bin, and each bin should have at least 3 observations (firms/individuals/households). This is required even for kernels with infinite support (e.g., Gaussian). For economic data, in addition to reporting the number of firms within each bin, you only need to provide concentration ratios (dependent on which dataset is used) for the full sample used to generate the kernel density plot. 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.
2. Cut off 5% from each tail. In doing so, we eliminate any 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. Chopping off 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 producing a KD plot of regression coefficients, you do not need to chop off the 5% tails. All relevant disclosure stats would need to be included in such cases.
3. Limit the detail of the scales on both axes. If values must be included, we suggest that the researchers use broad rounded numbers, not the values that are automatically generated by various software programs. In addition, the researcher needs to 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. If a researcher has significant concerns about degradation of data quality

using this approach, the researcher can 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.

To ensure that these rules are applied, researchers must create the kernel density inside the FSRDC.

II.D. Sign and Significance[1]

As mentioned in the Volume of Output section, the threat of disclosure risk from advanced attacks, like a database reconstruction attack, goes up considerably as more information is released. One way to minimize the volume of output is to only report the sign and significance of an estimate rather than the value of the estimate and its associated variance. Reporting sign and significance (sometimes abbreviated "S&S") can be an 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 can take less time than a traditional disclosure review.

To determine if delegated authority can be used to approve requested output or whether output needs to be reviewed by the DRB, sign and significance generally does not count towards volume of output. However, the DAO will use their judgement to determine if a large amount of sign and significance or a large amount of a combination of sign and significance and numeric estimates needs review 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. Implicit samples need to be tracked and appropriate disclosure statistics provided when any numeric estimates are being released for all samples contributing to the implicit sample. Please note that any sign & significance output that could pertain to Geographical Areas with Small Populations (GASPs) will still require population analysis – see Section IX. Releases Involving Geographical Areas with Small Populations (GASPs).

Here is an example. Say 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 are required:

1. The full disclosure statistics for A, B, and any implicit samples created by A and B
2. The explicit sample sizes of C and D

Note that in this scenario one would NOT need to worry about any implicit samples between A and C, A and D, B and C, and B and D since numeric estimates aren't being released for C and D. Further, any implicit sample between C and D also need not be considered for disclosure review. If later releasing numeric estimates for C and/or D, those full disclosure statistics and implicit sample considerations would become necessary for disclosure review.

[1] 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.

If the researcher adheres to 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 Administrator and DAO may determine on a case-by-case basis that additional disclosure statistics are required when sign and significance is reported. Note that for any prose or result summaries not framed solely around sign and/or statistical significance, researchers 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 it 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 can also be written in sentences. For example, "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. Additional support documentation is needed to properly perform the disclosure review - the researcher should provide all relevant output tables (not for release) along with the necessary disclosure statistics. The statements will also count toward the volume of output.

II.E. Programs

Most programs created in the FSRDCs may be released after being reviewed by CED-DA staff. 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 and 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.

If a program creates a subset of the data (e.g., by county or NAICS code), the researcher should be able to 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, and releasing programs showing this would not be permissible. 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 may not include non-public configuration information about Census Bureau computing systems. Programs may not include a "James Bond ID" for a Census Bureau data user, including the researcher.

Variable names and file layouts for Title 26 files are administratively sensitive and may not appear in programs. This is IRS policy. If you used a variable name that came directly from a Title 26 file, you must redact the variable name or replace it with a pseudonym that does not come from a protected file. With more journals requiring researchers to submit their code along with their research findings, CED-DA has seen a substantial increase in the number and size of code requests submitted for release. CED-DA urges you to modularize your code, creating macros or similar routines to do repetitive tasks rather than repeating nearly identical code several times in the program. This approach makes the code easier to review.

III. Samples and Implicit Samples

III.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 referred to as an analytical sample. For example, as a researcher you might define your sample as all firms in a dataset where firm age is five years or greater.

However, the set of observations in a regression model might only include firms where firm age is five years or greater and where there is non-missing data on all other variables used in the analysis. Disclosure statistics should be provided for the set of firms whose firm age is five years or greater and have non-missing data on all other variables used in the analysis. Similarly, the sample listed in your clearance request memo should be the set of firms whose firm age is five years or greater and have non-missing data on all other variables used in the regression analysis.

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, only include samples from a prior release if they are related to samples in the current requested output. For samples that evolve over different requests, a good practice is to use expanding naming convention so that you, the RDC Administrator, 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 request1111 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 – discussed below). **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 analysis at a level above individual or firm (e.g., county or industry), disclosure statistics should be calculated for the underlying sample of entities (either individuals or firms). For example, if estimating a coefficient for an industry level dummy variable, the disclosure stats would need to be shown for both the subsample of firms in-sample when dummy = 0 and the subsample of firms in-sample when dummy = 1. If using something like an industry level fixed effect and reporting a mean or percentile for those estimated coefficients, the disclosure stats should be generated for whatever underlying sample of firms was included in the calculation. In such an instance, reporting the mean would simply require the disclosure stats for all firms in-sample across all industries (or years, counties, etc.). If reporting the pseudo-percentile of such estimates, the disclosure statistics should be provided for the sample of entities underlying the pseudo-percentile calculation. If reporting an individual entity estimate (e.g., an industry maximum), the disclosure stats would pertain to the sample within that entity.

III.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. Such implicit samples must also be addressed in disclosure review and the usual disclosure criteria will apply to implicit samples.

One potentially helpful way to identify implicit samples is to consider whether knowing the sample sizes of two samples would yield the true difference between samples. For example, say there is Sample A containing 127 people and Sample B containing 75 people. If all 75 people in Sample B are in Sample A, then an implicit sample of 52 people appear in Sample A but not Sample B. If there are some people in Sample A but not in Sample B, and some people in Sample B but not 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 implicit samples relevant to their analysis, including those

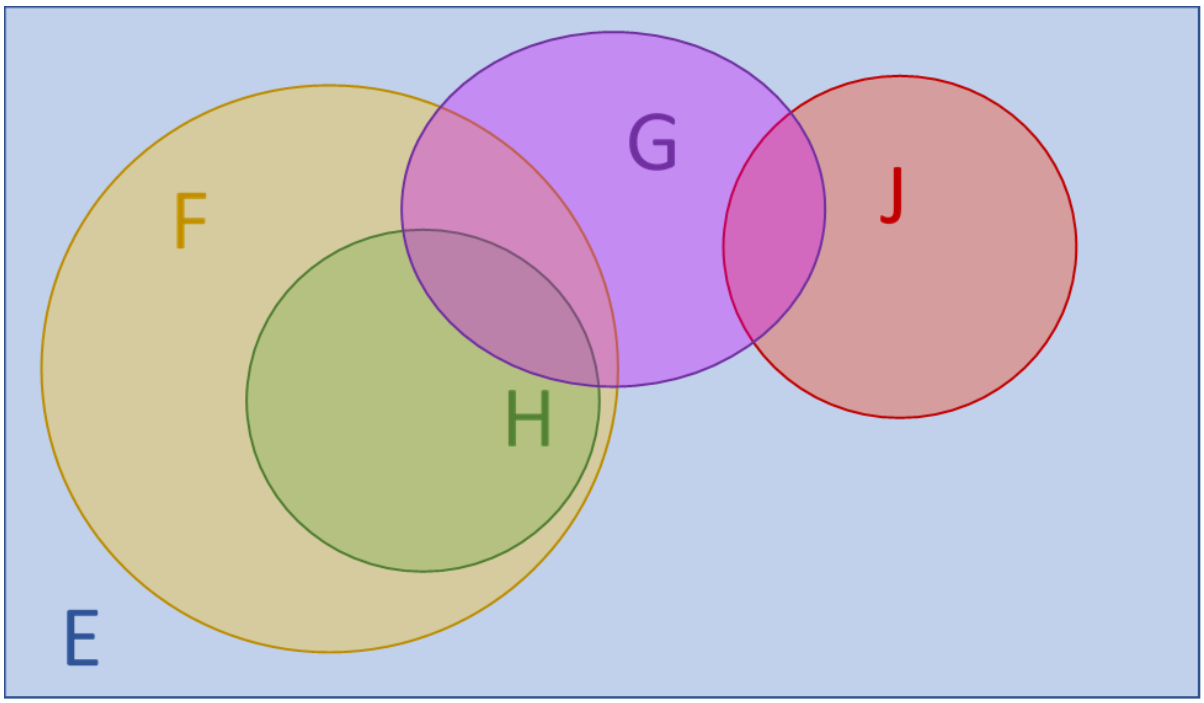
- Created from different sample definitions from which estimates, samples sizes, or other statistics are being released
- Within a certain release request
- Between a current request and a prior release
- Between the project and other published data, whether from standard publications, other FSRDC projects, 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 supporting files should show person counts for Sample A, Sample B, and Sample AB.
2. Mean employment was released for a sample of firms 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 Administrator 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.

As noted above, 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:

- **Problematic implicit cells are not created since the same estimates are not to be released for**

both sample and 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 implicit cells are not created because the estimates produced by the sample and sub-sample are not the same.

- Sign & Significance Output (see Section II.D. Sign and Significance)

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. Note that showing the disclosure statistics for the implicit sample of EFGJ (i.e., multi-unit, non-manufacturing firms outside the 10 most highly populated states firms) would be sufficient to cover Implicit Samples EF, EG, and EFJ; however, it is certainly fine to show the disclosure statistics for each separate implicit sample. 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 there is any uncertainty, please consult with your RDC Administrator.

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 an implicit sample making it impossible to release output from the new sample.

IV. Legacy Methods

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 be demonstrated to pass disclosure, as documented in supporting statistics demonstrating the adherence to all relevant disclosure rules. 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 implicit sample created. Below are the disclosure avoidance legacy methods.

IV.A. Cell Size Thresholds

The Census Bureau requires a minimum unweighted count for each cell. 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.** If the count is one, that particular unit, say a person, can easily find himself in that table. If the count is two, one of the units can find herself in that table, remove herself from the cell and perhaps identify the second person in that cell. She may know the person and know that they were in that same table cell for a given census or survey. Using this information, she may also be able to connect overlapping, related tables that can help connect a given person to that cell and discover additional information about that person. Another problem with small cells that appear in many different overlapping tables is that a data user could potentially link other information in those tables to form a microdata record for a small geographic area or even a partial microdata record. That record could be linked to a microdata file released from that same census or survey. The Census Bureau also requires microdata and tabular data (also known as magnitude 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 comingled data (from the Census Bureau and the IRS), IRS requires these thresholds:

<u>For establishment data:</u>
- At least 3 companies (firms) for national estimates.
- 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.

IV.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. Counts of highly aggregated entities like counties or industries do not need to be rounded assuming all disclosure checks pass for the underlying samples of individuals or firms.

Rounding is a legacy method that reduces information in a cell. Rounding rules applied to unweighted counts, weighted estimates, and model-based output have been revised to better protect 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.

IV.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 follows when statistical output and weighted estimates are reported in thousands, millions, or any other unit. Trailing zeros after a decimal place and beyond four significant digits are not allowed and will not pass disclosure review.

IV.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 from probability sample surveys (excluding unweighted counts), must be rounded to four significant digits as described above in order to be released. 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, you can work with your administrator to request an exemption from the DRB.

IV.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 reviews, please provide unweighted and unrounded counts in your support files. A good practice is to include the rounded and unrounded counts in your support file so that the RDC Administrator and DAO can make sure you've rounded correctly.

If you need to report unweighted counts for external release, 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.

IV.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. For example, if you are reporting the mean of binary variable 'sex' where 0=men and 1=women, you need to first round the number of men and women according to the unweighted count rounding rules and then take the ratio of the rounded values. If you have 485 men and 710 women, the rounded number of men is 500 and the rounded number of women

is 700. The reported mean would be ((500*0)+(700*1))/(500+700)=0.5833. The DRB has granted exceptions to this rule when the numerator and/or denominator of the percent or rate is not shown.

Alternatively, ratios can be calculated using unrounded numerators and/or denominators if you limit the precision of the resulting ratio. For thresholds based on a rounded, unweighted denominator (D) and an unrounded unweighted proportion (P):

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

When using this threshold, the reported numerators and denominators must be rounded. 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.

Note that for panel data, the "D" above pertains to the number of entities – not the total number of observations. For example, if there are 100,000 observations in the denominator but only 500 firms, the proportion P should be rounded to no more than 2 significant digits. In cases where it is ambiguous (e.g., when the traditional method would be 4 significant digits and the alternative would be rounded to 3 significant digits), researchers should indicate if they are using the traditional method.

IV.C. Concentration Rules

In economic data, the distributions of the variables are often highly skewed so that a few firms (even one) may account for most of the value in a cell; i.e., the cell is highly concentrated. We cannot allow the release of output that comes "too close" to revealing any individual firm's value. To make this practical, criteria are needed for deciding when releasing tabulated information in a cell comes "too close" to disclosing confidential information about an individual firm. Such a cell is called *sensitive*, and there are mathematical rules for determining which cells are sensitive. Two such rules are called the *p% rule* and the *(n,k) rule*. 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, a cell is sensitive if the second largest firm can determine the largest firm's value to within p%.

The value of the rule parameter p is confidential. The Census Bureau currently uses this rule in its publications. Under the (n,k) rule, which is used in Census Bureau publications 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 not confidential, but the value of k is confidential.

The FSRDCs use a combination of these rules; the general principle is to use the rules in effect for the particular survey and year in question. The general rules:

Standard Primary Disclosure Rules for Research Output based on Economic data
(from business establishments and firms)

The p% rule applies for 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 for 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 will use the (n,k) rule).

The values of p and k are confidential. Your RDCA will give you their values as needed since they may be revealed only to individuals with Special Sworn Status who need to know them.

The p% and (n,k) rules are conducted to protect against attribute disclosure of data for the most influential companies. The rules do not pertain to the name or demographics of that company; rather, these tests address sensitive information about the company (payroll, number employees, etc.). The statistics in question are not outliers but contribute greatly to the total for their industry or trade area.

Let $x_1, x_2, \ldots, x_N$ represent the contributions, in descending order, from respondents 1 through N with x_1 representing the largest contribution and X representing the sum of the N individual contributions.

The p-percent (p%) rule is integral in determining the sensitivity of a primary cell targeted for suppression. The p% rule takes the cell total minus the sum of the two largest contributions. This value must be at least p% of the largest contribution to pass disclosure review. Let X be the weighted aggregate total 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 safe:

$$|x_1| + |x_2| + \cdots + |x_n| \leq \frac{k}{100}X$$

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 something like employment should not be. When confirmed, their absolute values are used to run the p% and (n,k) rules.

For economic data prior to 1992, information product developers and researchers use the (n,k) rule; for data from 1992 and later, information product developers and researchers use the p%

rule to test output for possible risks of disclosure. The two rules are related. If the p% rule fails, then the (n,k) rule fails as well. If an aggregate passes the (n,k) rule, then it passes the p% rule when 100 – k = p.

On the other hand, passing the p% rule doesn't imply passing the (n,k) rule. In practice, reporting the results for both is best. The p and k values used at Census are highly sensitive. These values are never published and are considered confidential. Researchers will be provided these values by the administrator if/as needed.

The variables used in primary disclosure analysis (p%, (n,k)) differ across surveys, and within the survey they sometimes change over time. Table 1 and Table 2 list the key variables you should use to run disclosure analysis. Sometimes Table 1 and Table 2 may not be directly applicable to your dataset/year combination. The most important point is *tabular output for cells that the Census Bureau has suppressed will not be approved for release.* It is the responsibility of the researcher to check this. Please speak with an RDC administrator for guidance.

For analysis using data spanning multiple years (e.g., a pooled sample), if the key variable(s) change over time then use the most recent key variables associated with the most recent year of data in the analysis. If an analysis includes data from before 1992 and after 1992 then use p% as the measure of firm concentration.

If more than one survey is used, take a maximal approach in computing disclosure statistics for the key variables (e.g., if survey A uses payroll and employment as key variables and survey B uses employment and sales as key variables, then the key variables would be payroll, employment, and sales).

Concentration measures are calculated on the original variable as provided in the dataset, not on transformed measures, e.g., use total value of shipments and not ln(total value of shipments). If you take a continuous variable (e.g., investment) and recode it into a binary/categorical variable (e.g., a dummy indicating having positive investment or not) and the key variables are "all continuous variables used" then concentration ratios within the respective categories are required only for the underlying continuous variable (e.g., investment). In addition, concentration measures are calculated for key variables that are used to generate any other output variables, even if they themselves are not in the output. For example, many researchers use x, y, and z to generate Total Factor Productivity. They therefore must run disclosure analysis on x, y, and z.

Table 1 lists the key variables which require use of the p% or (n, k) rule whenever that particular dataset is utilized in an output request. For any Economic Census datasets (e.g., Census of Manufactures), please consult Table 2 which immediately follows Table 1. If you are using an economic dataset that does not have key variables listed in Table 1 or Table 2, calculate concentration ratios on all continuous variables used in your analysis. In the rare event that no key variable is listed and some continuous variables used in the analysis pass the disclosure thresholds while other continuous variables do not pass, the output will be reviewed by the DRB.

Disclosure statistics are required for all relevant samples (explicit and implicit) and subsamples. This

includes subsamples created by binary/categorical variables. If a dependent variable is continuous and a key variable (e.g., when the key variables are "all continuous variables used"), then concentration ratios must be calculated for that dependent variable by each category of binary/categorical independent variables. When the key variable is "all continuous variables used", if the dependent variable is categorical (e.g., logistic regression), then within each category of the dependent variable, concentration ratios must be calculated for any continuous variable that appears as an independent variable in such a regression. Concentration ratios are calculated at the firm level even if the analysis is performed at the establishment level.

If the requested output contains count measures on economic data (e.g., number of firms making investments each year) then only firm counts are required as part of the disclosure statistics for these measures. For non-count measures, firm counts and concentration ratios are required as part of the disclosure statistics.

Recall that the (n, k) rule is used for data before 1992 and the p% rule is used for data in 1992 and later. Further, the parameters p and k are confidential, and the values are only provided to researchers on a "need to know" basis.

Table 1: Key Variables Used for Concentration Ratios for Economic Data

Economic Dataset	***Key Variables for Disc. Statistics***	***Comments***
Annual Capital Expenditures Survey (ACES)	Every continuous variable used	Investment variables are not sufficiently correlated with one another to use a key variable.
Annual Retail Trade Survey (ARTS)	Every continuous variable used (previously Sales was key variable)	
Annual Business Survey (ABS)	None	Only firm counts are required (ABS has noise added).
Annual Survey of Entrepreneurs (ASE)	None	Only firm counts are required. Counts must meet IRS Pub1075 thresholds.
Annual Wholesale Trade Survey (AWTS)	Every continuous variable used (previously Sales was key variable)	
Business Expenditures Survey (BES)	Payroll	

Economic Dataset	***Key Variables for Disc. Statistics***	***Comments***
Business R&D and Innovation Survey (BRDIS), Survey of Industrial R&D (SIRD)	Every continuous variable used	Variables are not correlated with one another.
Business Register	Payroll and employment, Sales for Census years	
Census of Auxiliary Establishments (AUX)	Payroll and employment	
Commodity Flow Survey (CFS)	None	Only firm counts are required (CFS has noise added).
Current Industrial Reports (CIR)	Value of shipments	
Exporter Database	Every continuous variable used	No previous known key variables.
Foreign Trade Data – Exports (EXP), Imports (IMP)	Value of shipments (e.g. imports or exports) Total shipments	From 2005 on, this survey uses noise infusion in disclosure protection.
Integrated Longitudinal Business Database (ILBD)	Receipts	The noise factors need to be applied for tabular output but not regression output.
Kaufman Firm Survey	Every continuous variable used	No previous known key variables.
Longitudinal Business Database (LBD)	Payroll and employment	
Longitudinal Employer-Household Dynamics (LEHD) Infrastructure Files	Employment or Payroll (if all released variables are not at the individual level)	See table in Subsection IV.C.1. Special Case: Person-Level Analysis using Firm Data to determine what disclosure statistics are required for person-level analysis.

Economic Dataset	*Key Variables for Disc. Statistics*	*Comments*
Longitudinal Firm Trade Transactions Database (LFTTD)	Imports or exports	A "total value of shipments" is not available in the LFTTD. EXP and IMP don't have the usual firm identifiers. Use the closest proxy instead. For EXP, use EIN if available, or SSN/PIK or CBP otherwise. For IMP, use EIN if available, or name values otherwise
Manufacturing Energy Consumption Survey (MECS)	Every continuous variable used	Variables are not correlated with one another.
Manufacturers' Shipments, Inventories and Orders (M-3)	Value of shipments	
Manufacturers' Unfilled Orders Survey	Every continuous variable used	No previous known key variables.
Medical Expenditure Panel Survey-Insurance Component (MEPS-IC)	Total employment	
Management and Organizational Practices Survey (MOPS)	Every continuous variable used	This is a supplement to the ASM.
Monthly Retail Trade Survey	Every continuous variable used	No previous known key variables.
Monthly Wholesale Trade Survey	Every continuous variable used	No previous known key variables.
National Employer Survey (NES)	Plant count (continuous) and smaller plant count (discrete)	This survey is less sensitive than other surveys because of no certainty strata, limited geography, and use of variables with discrete responses.
Ownership Change Database	Every continuous variable used	No previous known key variables.
Quarterly Financial Reports (QFR/QFRCEN)	Every continuous variable used	

Economic Dataset	*Key Variables for Disc. Statistics*	*Comments*
Quarterly Services Survey	Every continuous variable used	No previous known key variables.
Quarterly Survey of Plant Capacity Utilization (QPC/PCU)	Every continuous variable used	
Services Annual Survey (SAS)	Every continuous variable used	
Survey of Pollution Abatement Costs and Expenditures (PACE)	Every continuous variable used	Variables are not correlated with one another.
Standard Statistical Establishment List (SSEL)	Payroll and employment	
Survey of Business Owners (SBO)	None	Only firm counts are required.
Survey of Manufacturing Technology (SMT)	None	Only firm counts are required.

Table 2: Key Variables Used for Concentration Ratios for Economic Census Data

Economic Census Dataset	*Key Variables for Disc. Statistics*	*Comments*
Census of Manufactures & Annual Survey of Manufactures (CMF/ASM, formerly LRD)	Value of Shipments and Capital Investment (pre-2004); every continuous variable used (2004 onward)	(1) If using pre-2004 data, use the TVS variable here; if TVS is not included in the data file, then shipments or payroll can be alternatively used to meet this requirement. (2) These surveys use an unweighted numerator and weighted denominator to calculate shares.

Economic Census Dataset	*Key Variables for Disc. Statistics*	*Comments*
Census of Construction Industries (CCN)	Value of shipments (pre-2002); every continuous variable used (2002 onward)	
Census of Finance, Insurance, and Real Estate (CFI),	Value of shipments (pre-2017); value of shipments and employee payroll (2017 onward)	
Census of Mining (CMI)	Value of shipments and capital investments (pre-2007); every continuous variable used (2007 onward)	
Census of Retail Trade (CRT)	Value of shipments (pre-2017); value of shipments and employee payroll (2017 onward)	
Census of Services (CSR)	Value of shipments (pre-2017); value of shipments and employee payroll (2017 onward)	
Census of Transportation, Communications, and Utilities (CUT)	Value of shipments (pre-2017); value of shipments and employee payroll (2017 onward)	
Census of Wholesale Trade (CWH)	Value of shipments (pre-2017); value of shipments and employee payroll (2017 onward)	
Census of Island Areas	Value of shipments (pre-2017); no key variables after 2007 (noise added)	

IV.C.1. Special Case: Person-Level Analysis using Firm Data

There is a special case to consider whenever you are using linked person-employer data and performing analysis on persons rather than firms or establishments. Consider the possibility that 950 individuals in a sample of 1000 workers were employed by one firm with exactly 950 workers total. The remaining 50 individuals in the worker sample came from 50 different firms and each of these firms also had close to 1000 workers total. If we treated the set of employers of this sample of people as a sample of firms and calculated a traditional concentration ratio, we would not detect any problems. The firms are close to each other in total size and, measured this way, one large firm will not dominate the sample. However, the

person-level statistics we released would be heavily dominated by the employees of one firm and could release the entire earnings distribution for that firm.

If a sample of firms is selected to study the workers at those firms (or establishments) in relation to some type of firm (or establishment) outcome variable, then traditional concentration ratios would still be 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
- Concentration ratios based on the number of workers in the sample employed by each firm (workers-in-sample concentration ratios)

These concentration ratios would effectively describe the concentration of workers at the employers in the sample or, thinking from the other perspective, how many firms there were relative to people. The disclosure concern in such special cases is that the statistics on such samples of workers could pose a disclosure risk if the workers were highly concentrated at a small number of firms.

The relevant disclosure statistics required for different types of analysis are summarized below:

Unit of analysis	**Is firm/establishment data used?**	**Required Disclosure Statistics**
Firm/Establishment	YES	Firm counts Traditional concentration ratios
Person	YES	Person counts Workers-in-sample concentration ratios (see above paragraphs)
Person	NO	Person counts

V. 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 Administrator 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. 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.

VI. Reporting Match Rates

VI.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%. 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 to no more than four significant digits.

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.

VI.B. Protected Identification Key (PIK) Rates

PIK rates do not need to be top-coded but must be rounded to no more than four significant digits.

VII. Special Considerations for Particular Surveys

VII.A. American Community Survey

The ACS is not intended 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. It may be barred for particular projects if specific concerns arise during proposal or output reviews.

The following is generally allowed:

- **Block-level contextual variables in an individual- 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.

Both ACS and decennial block-level contextual variables are allowed to be included in models using ACS data. Using blocks in this way will typically not present a problem if the analytical sample consists of households from the entire nation. However, samples for lower levels of geography may introduce concerns.

Summary statistics and tables describing the analytical sample will be carefully considered, as described above.

As usual for disclosure review of models, cell counts for any dummy variables must meet a threshold. Further, 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.

VII.B. Longitudinal Employer-Household Dynamics

For the 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 IV.C. Concentration Rules for more information. Here are some scenarios and the disclosure statistics required in each:

- If person level analysis and no firm/establishment data is being used, only counts would be required
- If person level analysis and firm/establishment data is being used, person counts and the workers-in-sample concentration ratios would be required (see Subsection IV.C.1. Special Case: Person-Level Analysis using Firm Data).
- If firm or establishment level analysis using T13 components of LEHD only, firm disclosure stats should be calculated using SEIN.
- If firm or establishment level analysis using T26 components of LEHD, firm disclosure stats 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.[2]

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 Administrator about these requirements and plan for any needed extra review time.

The LEHD is subject to other special rules. Please see the three-page memo on "LEHD Disclosure Avoidance Review Protocols," revised April 30, 2013.

VII.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).

[2] 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.

VII.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. 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. Here is an 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

VIII. 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, the Disclosure Review Board (DRB) must deliberate about items with a large amount of output. 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 a given request reduces the likelihood of approval. This is particularly true as researchers produce additional requests in the same line of research.

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 released an estimate and later re-generate the same estimate after correcting a coding error, using different weights, etc. Releasing the second estimate would still count toward volume of output. Further, in unusual cases, updating previously released numeric estimates could cause more direct disclosure risks unrelated to volume of output concerns.

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 and indicate any relationships (or a lack thereof) between samples within or across requests. We understand it is 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 RDC projects. When possible, researchers should be aware of other research being performed on related data samples under the same FSRDC project.

Two specific limits have been set regarding the number of estimates that can be released by a DAO using

their delegated authority without explicit DRB approval. These two limits are not exhaustive. On unusual occasions, an RDC Administrator or DAO may recommend that because of a request's particular features the request should be taken to the DRB because of the volume of output. In particular, the DRB will likely be asked to make the final approval if a high number of estimates is being generated for a small sample of entities even if neither volume of output limit is technically violated.

VIII.A. The Absolute Number of Estimates Cap

The DRB has set an absolute limit of 5,000 estimates for delegated authority. This cap is cumulative **across related requests**. The DAO can exercise delegated authority on output amounts up to 5,000 estimates; however, any summed volume of output in excess of 5,000 estimates will automatically require DRB review.

VIII.B. The 30:1 Ratio Rule

Output requires DRB approval if the ratio of the unweighted sample size to the total number of estimates is less than 30:1. The "sample size" is for the main number of observations (usually the unweighted entity count) used for the analysis. For example:

- If Samples 2 and 3 are subsets of Sample 1, the 30:1 ratio could be the entity count in Sample 1 divided by the total number of estimates requested for any analysis on Samples 1, 2, and 3, or it may be based on each sample separately.
    - ➔ Say that the research for Request 1 pertains to multi-unit firms and then there is supplemental analysis on multi-unit firms in the manufacturing industry and multi-unit firms in the services industry. The sample size would be total number of multi-unit firms, and the number of estimates would be the sum of estimates requested for release across analyses for Request 1.
    - ➔ If Sample 1 is identified as a distinct research sample (i.e., not just "the entire ACS") but is **not** used for analysis, the 30:1 ratio calculation will depend on the relationships among samples. For example, if Sample 2 and Sample 3 are disjoint (e.g., men and women in the ACS), then separate 30:1 ratio rule calculations would be applied. If there is overlap between Samples 2 and 3, the 30:1 ratio rule would typically treat Sample 1 as a "parent" sample.
    - ➔ If Sample 1 is from dataset A and Sample 2, a subset of Sample 1, merges on information from dataset B, 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 on the estimates derived directly or indirectly from Sample D, with another 30:1 ratio calculated for Sample E and its associated estimates.

- For panel surveys where the output is not reported by year, the entity is the person, household, or firm, not crossed by year. For example, if your sample is of 1,000 unweighted individuals and individuals respond once a year for five years, the sample size remains 1,000. If reporting estimates by year, the 30:1 ratio will 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 individuals, but there are 100 regression estimates generated from a subsample of Sample 1 consisting of only 500 individuals, such a request would require the DRB for final 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 should recognize when they are producing a high number of estimates from a small sample of entities even if the main sample is large. Researchers working with small analytical samples and/or requesting a large volume of output should aim to have their request ready for review several weeks before any researcher deadlines to allow for enough time to take the request to the DRB.

Attempts to circumvent these rules by splitting up output requests, adding samples, etc. are strongly discouraged. Such requests would be treated as a single request and taken before the DRB for review. Going to the DRB does not mean that your output will be rejected, but simply that the DAO cannot use their delegated authority to approve the requested output above the volume of output limits. If the RDC Administrator or DAO determines that researchers on a project are trying to "game the system", they may recommend those requests be rejected outright.

The following are possible alternatives that researchers can consider to limit the number of estimates requested:

- Consider not reporting the coefficients on less important variables (e.g., geographic or sector-specific dummy variables, or fixed effects) in a model when reporting coefficients on the variables of interest.
- Consider combining dummy variables.
- Instead of reporting parameter estimates, consider reporting the sign of the estimate and its significance level.

VIII.C. What counts as an estimate?

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

Note that the sample size or number of observations only counts 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.

Graphical output and figures count toward both volume of output thresholds described earlier. 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) would count as 40 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. Figures may be viewed more favorably from a disclosure risk perspective if the resolution of the figure reduces the precision of the estimates.

Examples of volume of output calculations are included in Appendix B: Volume of Output Examples of this handbook.

IX. 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 increased re-identification risk that accompanies database reconstruction and record linkage from external data. Geographical Areas with Small Populations have been determined to have a higher disclosure risk, and thus have become first in line to require noise injection. Note that GASPs used to be known as "substate" in previous versions of disclosure guidance.

IX.A. Definition of a GASP

A GASP is defined as the union of one or more part-state geographies with a 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, counties, or tracts). Aggregates of small areas collectively larger in population than the least populous U.S. State are not considered GASP. This definition has two key components:

GA } Geographic selection of a part-state geography for analysis

SP } Selected geography is less populous than the least populous contemporaneous state

An example of a potential GASP would be 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 GASP. See **Figure XI.A.1** below. The counties in this geographic region have a small population.

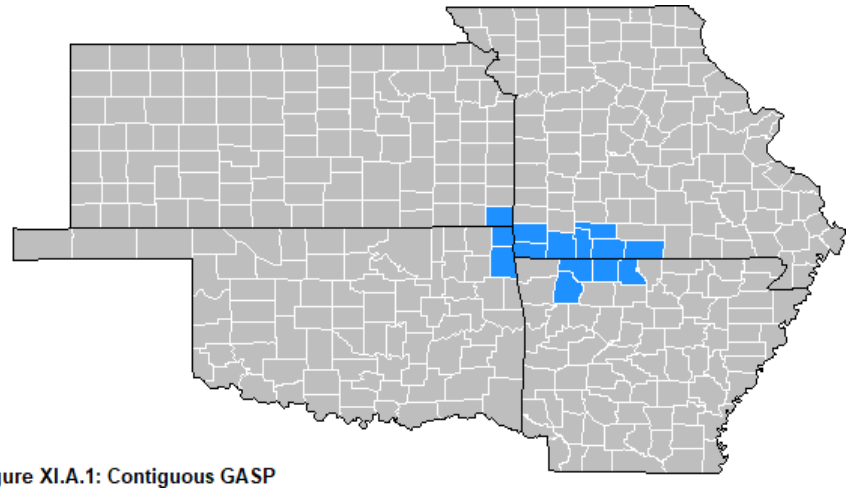


Figure XI.A.1: Contiguous GASP

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

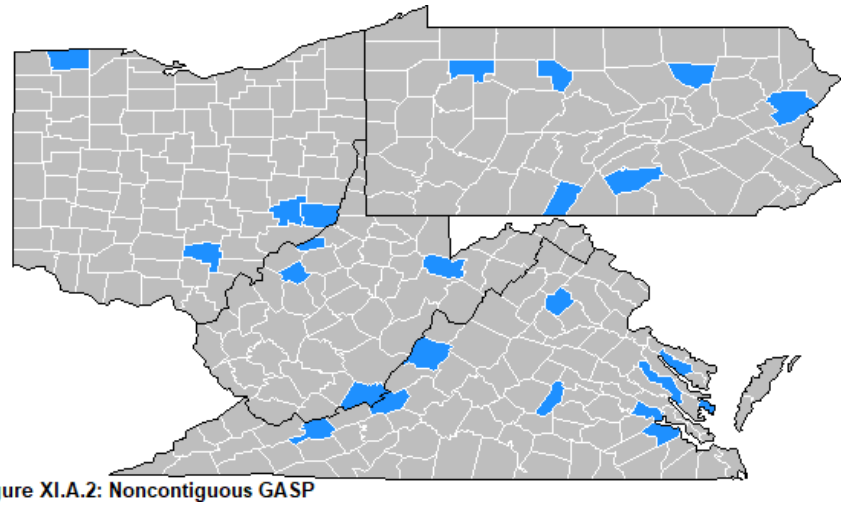


Figure XI.A.2: Noncontiguous GASP

These criteria must also hold for implicit samples. That is, geographic slivers that fall below the population threshold when subtracted from a parent sample are also considered GASP. Consider the following example:

> **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) could be considered a GASP if the same analysis is done on samples A and B.

IX.B. Population Analysis

Population analysis refers to a comparison of a geography's population estimate to the population of the least populous state for the same period. This is done by summing the populations of the geographical areas used in the proposed output and comparing the sum to 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 would not need population analysis if the sample is national. Similarly, a regression consisting of only continuous variables at the tract-level would 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 due to publicly known changes in the part-state geographies in sample. Effectively, if the sample of geographic entities is known or being made known, then population analysis is required. If the geographic entities in sample are not known 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 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. Here are 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 where the cities are explicitly listed (e.g., our sample contains observations from Cities A, B, C, and D).

Cases where population analysis is likely not required would 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.

Note that 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 still be performed even if the definitions are made based on internal data only.

Geographically defined regression categories may require additional population analysis. This includes:

- Geographically defined categorical variables included by themselves or crossed with continuous variables.
- An interaction term involving more than one categorical variable, with at least one 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 only required 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). Say that these block-level variables are all based on internal data. Population analysis would then only be required for the collection of cities.

Note that 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 all of those years.

The population analysis will either pass or fail. Products that pass the population analysis can be released directly by a Disclosure Avoidance Officer (DAO). Products that do not pass population analysis will need to go before the Disclosure Review Board (DRB) and will likely require noise injection or another approved methodology.

IX.C. Model-Based Estimates

A model-based estimate is defined here as those 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. Potential 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 which 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
- Model-based estimates that have DRB recommendation for approval, and have been approved by the Chief Scientist and DSEP

To determine if a model-based estimate does not require noise injection, a detailed description of the model is needed to demonstrate it fits the conditions. This 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
- 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 can be released

- Multilevel Regression and Post-stratification (MRP) models

IX.D. Noise Injection

If the above conditions for direct and model-based estimates are not met for GASP, estimates for sub-national geographies will only be released by the DRB 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 infused 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.

CED-DA will work with you on the noise injection process. A consultation with the Center for Enterprise Dissemination-Disclosure Avoidance (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.

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 initial work on this approach to semantics) and

  https://arxiv.org/abs/1905.02383

XI. Resources

There are a number of resources available to help researchers follow disclosure avoidance guidance and generate disclosure statistics.

XI.A. Practical Applications of Disclosure Rules

Rounding
A rounding program to check spreadsheets is available to researchers on the IRE. The program and instructions can be found in /data/support/researcher/programs/rounding. The rounding program will catch most rounding errors. However, this rounding program is not 100% effective at catching rounding issues and will sometimes indicate an estimate should be rounded to the count rule. We encourage you to review the results of the program prior to submission.

Concentration Ratios
There is a Stata program available to researchers working on economic data to help prepare disclosure statistics, including concentration ratios and sample counts. The file is here: /data/support/admin/disclosure/discstats.do. And a tutorial of how to use the program can be found here: /data/support/admin/disclosure/DisclosurePrimer_Stata_oct16.odt.

Population Tables and Population Analysis Tutorial
For output involving Geographical Areas with Small Populations (GASPs), a population analysis is required to determine if the population of the geographical area is smaller or greater than the least populous state. To assist with the population analysis, some population tables of counties and MSAs for various years have been uploaded to the IRE. Your RDC Administrator will help you get access to these data tables. These tables cover many of the GASPs used by researchers, but you might need to use other resources to do your population analysis. Work with your RDC Administrator to ensure you are conducting your population analysis correctly. A population analysis tutorial is forthcoming.

XI.B. Disclosure Tutorial and Request Memo Examples

Disclosure Tutorial (Forthcoming)
A disclosure tutorial is being created using the Survey of Business Owners (SBO) public use data. The tutorial includes Stata programs that clean data, run an analysis, and run disclosure statistics. The tutorial also has a Clearance Request Memo to show researchers what information to include in the memo and what to add to output folders and support folders. Ask your RDC Administrator where to find the tutorial. The tutorial can be run inside or outside of the FSRDC.

Clearance Request Memo Examples
RDC Administrators have pulled examples of clearance request memos for you to review and use as guides on how to write a clean, clear, and concise Clearance Request Memo. Ask your RDC Administrator where to find these examples.

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.

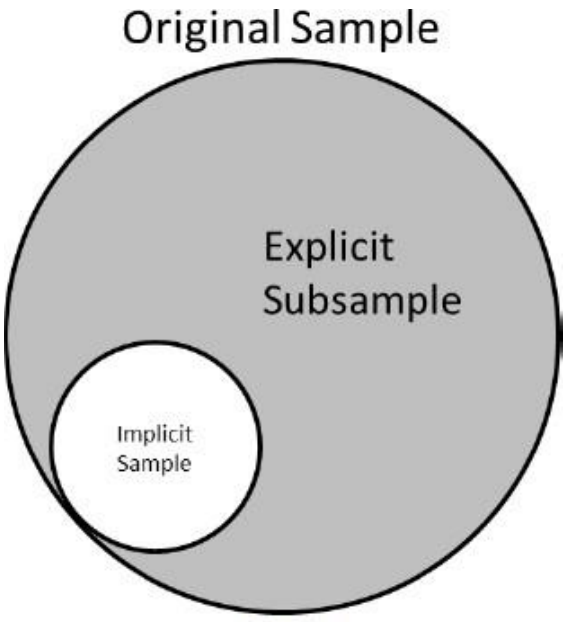


Figure 1

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 Administrators 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.

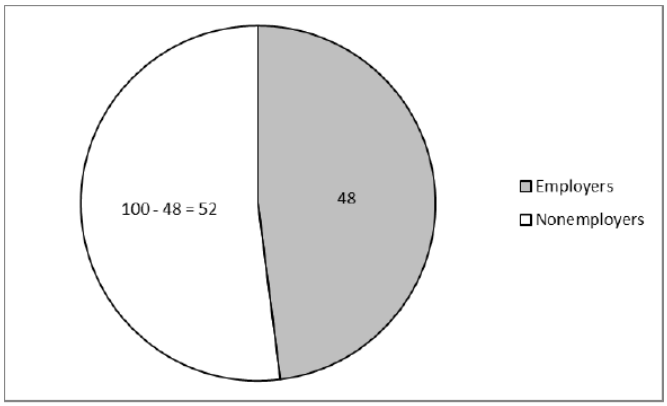


Figure 2

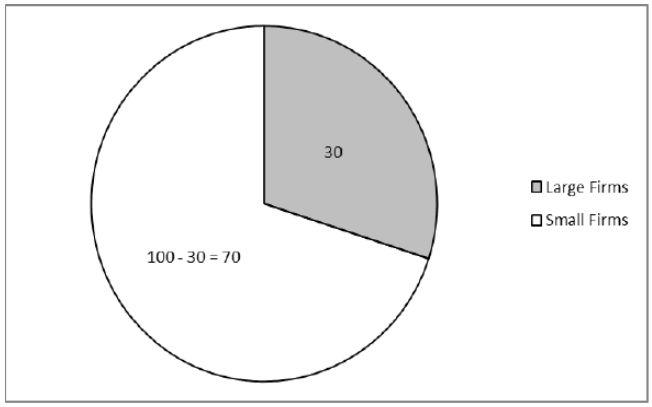


Figure 3

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.

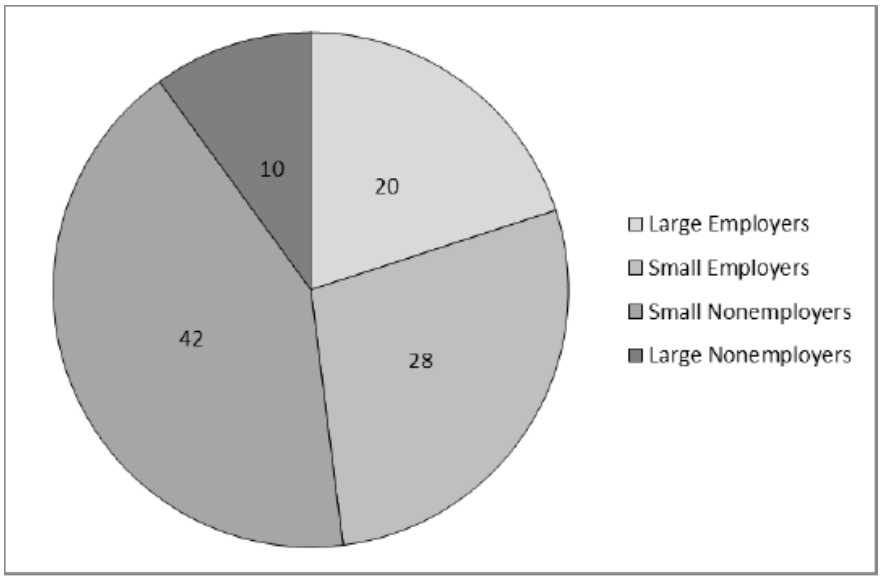


Figure 4

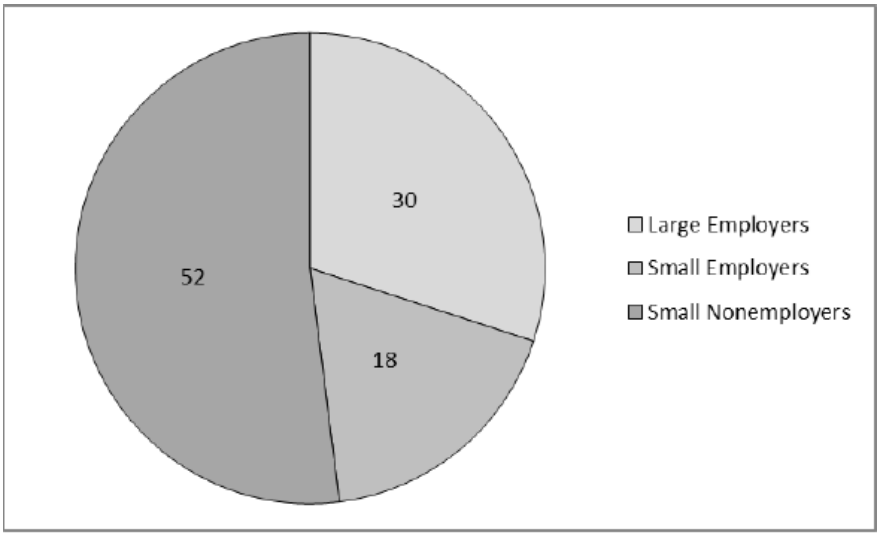


Figure 5

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.

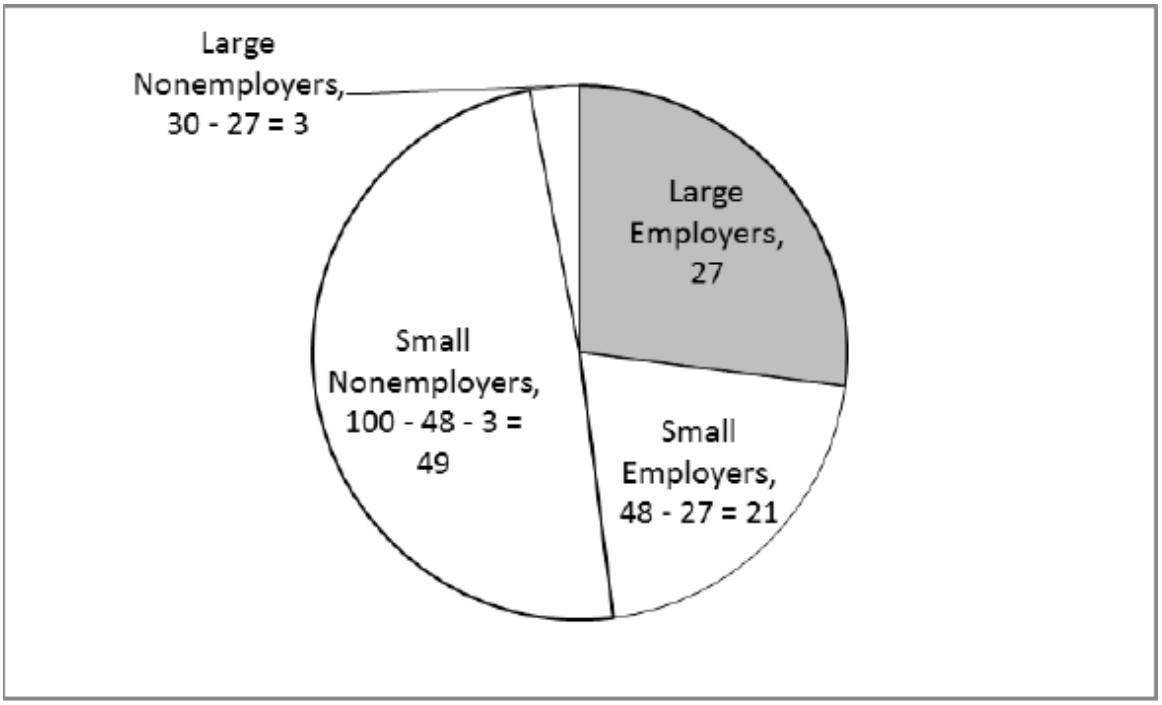


Figure 6

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

All of the examples below stem from 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).

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 'Y' (yes) indicator for fixed effects and the 'NR' (not reported) cells do not count. Also note that the number of observations across the tables/models are the same so they only count once.