Thomas Stratmann
Mercatus Center Scholar
Professor of Economics, George Mason University

to

Representative Scott Garrett
Chairman of the Capital Markets and Government Sponsored Enterprises Subcommittee
US House of Representatives
2232 Rayburn House Office Building
Washington, DC 20515

January 23, 2014

Dear Chairman Garrett:

Thank you for this opportunity to comment on the necessity of the scope of data collection of sensitive financial information by the Consumer Financial Protection Bureau (CFPB). I believe that the CFPB is collecting far more data than necessary. This expansive data collection is both expensive and risky. As will be demonstrated, a one percent sample will achieve the CFPB's goals while alleviating concerns about consumer privacy and costs.

The CFPB's Current Practice

The CFPB has been collecting individual loan and credit card data from major US banks as part of its authorization under the Dodd-Frank Wall Street Reform and Consumer Protection Act (Dodd-Frank Act). The letter of request from the House Committee on Financial Services dated January 22, 2014 states:

> According to the CFPB, the combined data collected from the eighteen card issuers represent approximately 85–90 % of the outstanding card accounts. The U.S. Census Bureau projects that there were approximately 1.167 billion credit cards in the United States held by 156 million card holders in 2012.[1] Accordingly, the CFPB appears to be

[1] US Census Bureau, *Statistical Abstract of the United States: 2012* (Washington, DC: 2011), Table 1188, http://www.census.gov/prod/2011pubs/12statab/banking.pdf.

collecting account-level data on at least 991 million credit card accounts, which would correspond to roughly 60% of the adult U.S. population.

According to a CFPB request for proposals, "Account-level information provides unique insight into understanding changes in the credit card market. [. . .] Such information maintained in a database can be used to create both present-day snapshots and historical trend data and help the CFPB understand the cost of credit and how the costs are realized by consumers."[2]

It is my opinion that the CFPB is collecting much more data than necessary to conduct a valid statistical analysis of consumer financial markets. There are costs and potential harms to collecting and maintaining massive, comprehensive databases of personal financial information; these include storage and transmission requirements, potential for abuse or violation of consumer privacy, and security concerns in the event of a data breach. These costs and potential harms can be significantly reduced by using sampling methods to conduct an analysis of these data.

Sampling Techniques

Sampling involves collecting data for random smaller subsets of individuals instead of collecting data for the entire population. CFPB researchers can use the averages from these subsets—along with some aggregates reported from the banks—to create valid estimates for all the variables currently being used while collecting far fewer individual accounts' data.

Almost all of the data referred to in a CFPB example report from the month of September (attachment 8) are totals (counts and sums), averages, or percentages.[3] Counts and sums include the number of total accounts, the number of active accounts, and totals for commitments and outstanding loans. One cannot determine the total number of accounts, or total credit outstanding from information about a subset, but these totals could be easily reported as totals and so do not require granular data. The descriptive statistics, such as percent of balances 30+ days delinquent, average credit line, average original FICO score, etc., can all be accurately estimated from samples. The CFPB could use much smaller samples to estimate averages that would still be very precise.

In general, when analyzing averages and percentages the average of a subsample can be a very good estimate for the actual average in the population. With a large enough subsample, the expected error in estimates can be brought within any predefined tolerance for error. With the information the CFPB has already collected, researchers at the CFPB can easily determine how

---

[2] Consumer Financial Protection Bureau, *Request for Proposals: RFP # CFP-12-R-00001, Collection, Transmission, Validation, Aggregation, Reporting, Storage, and Analysis of Credit Card Data (CCD Services)* (Washington, DC: January 27, 2012), 5, https://www.fbo.gov/index?s=opportunity&mode=form&tab=core&id=61f9e255acb3ac044ffeb4ae10c6ec00.

[3] CFPB, *Collection, Transmission, Validation, Aggregation, Reporting, Storage, and Analysis of Credit Card Data (CCD Services), Amendment 1, Attachment 8*, July 14, 2011, https://www.fbo.gov/utils/view?id=00c122f39215846c6512612f816d749f.

large is a "large enough" sample size using the standard deviation and tolerance for error of each variable.

The term *standard deviation* describes a commonly used statistic that indicates how "spread out" the data is relative to its average value. The standard deviation is calculated routinely from any set of numbers. The term *tolerance* describes something a little more nuanced than a simple formula, and the value of the tolerance used is context dependent. Tolerance is used in experimental statistics where one conducts "power-analysis" before deciding how many subjects to enroll (and pay for). If one has a treatment that one thinks will increase a variable by some amount, power analysis looks at how likely one is to find statistically significant differences from the null hypothesis for different hypothetical "true values" of that variable for a given sample size. The key is to figure out how small of an effect one wants to be able to reliably detect—with that information, one can fairly easily determine how large is "large enough."

For an example in the matter at hand, consider the "average balance per account" variable. If CFPB researchers are using this variable to inform their analysis, then there is a level of tolerable imprecision that still allows for a valid statistical analysis. That is, if the actual average balance for some subset of accounts is $3,000, then it probably does not drastically alter research findings or policy recommendations if statistical sampling of a smaller subset yields an estimate of $3,001 or even (probably) $3,010. But it is easy to see how estimates that are off by $500 or some other large amount could negatively impact the bureau's ability to perform research and monitor credit markets.

If the bureau switched to statistical sampling to gather its data, researchers could determine the necessary sample size by fixing a tolerance (e.g., not wanting estimates to be off by more than $100 for 95 percent of the time) and applying some calculations based on the standard deviations in their existing data. If the standard deviation was usually $1,000 (i.e., at least 75 percent of accounts have balances within $2,000 of the average account),[4] then samples of 400 random accounts per subgroup would be sufficient for estimates that meet the required tolerance based on common, reasonable statistical assumptions.[5,6]

---

[4] Per Chebyshev's inequality, which states that at least $1 - \frac{1}{x^2}$ of any distribution will be within x standard deviations of its mean.

[5] Specifically, the Central Limit Theorem, as discussed on page 29 in George E. P. Box, J. Stuart Hunter, and William G. Hunter, *Statistics for Experimenters: Design, Innovation, and Discovery*, 2nd ed. (Hoboken, NJ: Wiley-Interscience, 2005).

[6] These numbers come from calculating $\text{standard error} = \frac{\text{standard deviation}}{\sqrt{\text{sample size}}}$ and assuming that the distribution of sample means will be approximately normally distributed about a population mean. I determine 95 percent or 99 percent confidence intervals as +/- the standard error times two or three, respectively.

While I chose the numbers in the example above for their simplicity, they reflect the ease with which sample sizes can be determined by tolerance for error and standard deviation. The general rule of thumb is $\min(sample\ size) = (2 * \frac{standard\ deviation}{tolerance})^2$ for 95 percent confidence intervals, or $\min(sample\ size) = (3 * \frac{standard\ deviation}{tolerance})^2$ for 99 percent confidence intervals. Presumably, the CFPB could set its cohort sample sizes based on the variables with the highest standard deviation and lowest tolerance for imprecise estimates.

Although I do not have access to data the CFPB collected, I can draw some inferences regarding the maximum number of data points that have to be collected, based on worst-case scenario estimates. Many of the variables in the example September document (attachment 8) are reported as percentages. These are convenient variables for my estimation, because for percentages, the maximal variance is 0.25,[7] so the maximal[8] standard deviation is 0.5. With only 40,000 observations, the 95 percent confidence interval is approximately +/- 0.005 (half a percent), and even the 99 percent CI is less than +/- 0.0075.[9]

Therefore, if the CFPB researchers decide their estimates of percentage variables need to be within one percent of the true value at least 99 percent of the time, then that would be achievable with sample sizes of 40,000 per subgroup of consumers.

The example report from September shows accounts broken up by FICO score (10 categories), origination channel (7 categories), bank and risk profile (9 categories each). Even if the CFPB were treating each of these categories as independent and drawing 40,000 new observations per category, that would still only require collecting data for 1.4 million accounts for the 35 divisions (the sum of subcategories in the categories "Mix by Origination Channel," "Mix by Refreshed FICO Score," "Bank Profile," and "Risk Profile" in Attachment 8). This number of 1.4 million accounts is well short of the reported 991 million accounts for which they are currently collecting data. If one were to collect data from 1.4 million individuals, instead of accounts, then these 1.4 million observations would be approximately one percent of the credit card holding public.

[7] Because percentages are bounded from 0 to 1.

[8] The standard deviations will nearly always be lower if the observation-level variable can take values besides 0 or 1 (e.g., percent of total unpaid balance) as opposed to variables like percent of full pay accounts. But somewhat more importantly, most percentages (e.g., percent of accounts that pay in full, percent of balances over limit) should be easily obtainable from the banks without requiring granular aggregation at the CFPB.

[9] $95\%\text{Conf. Int.} \cong \pm 2 * std.err = \pm 2 * \frac{.5}{\sqrt{40{,}000}}$ and $99\%\text{Conf. Int.} \cong \pm 2 * std.err = \pm 3 * \frac{.5}{\sqrt{40{,}000}}$.

Conclusion

Limiting their sampling to one percent of the relevant population would bring CFPB more in line with the US Census Bureau, which makes anonymized granular data available to researchers through the Public Use Microdata Sample (PUMS) and only provides one percent and five percent samples to researchers for statistical analysis. I see no a priori reason to think that credit data are any different than data collected by the Census, in terms of means relative to variance, so collecting a much smaller credit card sample should suffice.[10] Because of these factors, I believe that the CFPB should be able to conduct its research with data sampling, which may alleviate some of the concerns about cost and consumer privacy.

Sincerely,

Thomas Stratmann

[10] Additionally, the large-scale data that has been collected so far gives the CFPB anchoring values to ensure that sampling is giving them reasonable estimates. If initial estimates of averages (from smaller samples) are way off from the previous (near total) population averages, that would let the CFPB know which parts of the sampling procedure may need to be tweaked. Presumably, this is similar to how the Census Bureau uses the decennial census to complement and calibrate their survey sampling.