8

Sampling and Massive Data

Sampling is the process of collecting some data when collecting it all or analyzing it all is unreasonable. Before addressing why sampling still matters when massive amounts of data are available and what the new challenges are when the amount of data is massive, an overview of statistical sampling is provided below.

COMMON TECHNIQUES OF STATISTICAL SAMPLING

Random Sampling

In simple random sampling, everything of interest that could be sampled is equally likely to be included in the sample. If people are to be

sampled, then everyone in the population of interest has the same chance to be in the sample. For that reason, a simple random sample gives an unbiased representation of the population.

Simple random sampling is usually straightforward to implement, but other kinds of sampling can give better estimates. Stratified random sampling partitions the population into groups called strata and then randomly samples within each group. If strata are less diverse than the population as a whole, then combining group-level estimates, such as group means, is better than estimating from the population without partitioning. The improvement due to partitioning is larger if the more heterogeneous groups are more heavily sampled. Often the groups and group sizes are chosen to optimize some criterion—such as constraining the mean squared error of an estimate—based on past data, theory about the processes generating the data, or intuition. Of course, a sampling strategy that is optimal for estimating one kind of parameter may be inefficient or even unusable for another, so there is often an informal compromise between optimal stratification and pure random sampling so that useful information can be obtained about a wider set of parameters. This kind of compromise is especially important when data are expensive to collect and may be reused for unforeseen purposes.

There are many variants of random sampling beyond stratification. For example, random sampling may be applied in stages. For example, cluster or hierarchical sampling first randomly samples city blocks, apartment buildings, households, or other "clusters," and then randomly samples individuals within each sampled cluster. Panels that monitor changes in attitudes or health track a random sample of respondents over time, removing respondents from the panel when they have served for a fixed length of time, and replacing them with a new random sample of respondents. Only a fraction of the respondents enter or leave the panel at the same time, so the panel is always a mix of different "waves" of respondents. Nielsen runs a panel of households for television viewing, Comscore has a panel for Web browsing, and the U.S. government has many panels, such as the National Longitudinal Studies run by the Bureau of Labor Statistics. Other examples of staged sampling can be found in

environmental science and ecology. For example, points or small regions are randomly sampled, and then data are taken along random directions from a landmark, such as the center of the sampled region. Of course, random sampling methods can be combined, so panels can be selected with stratification, for example.

Sampling may also be adaptive or sequential, so that the sampling rule changes according to a function of the observations taken so far. The goal is to over-sample regions with "interesting" data or more-variable data to ensure that the estimates in those regions are reliable. For example, the number of observations taken in a region may not be fixed in advance but instead depend on the data already observed in the region or its neighbors. In sequential sampling, the population to be sampled (called the *sampling frame*) may change over time, as happens when sampling a data stream.

Randomly sampling a data stream may mean producing a set of observations such that, at any time, all observations that have occurred so far are equally likely to appear in the sample. Vitter (1985) called this reservoir sampling, although his algorithm was earlier known as the Fisher-Yates shuffling algorithm (Fisher and Yates, 1938). The basic idea behind reservoir sampling is that the probability that the next item seen is sampled depends on the number of items seen so far, but not otherwise on their values.

Random sampling of data streams has two major disadvantages: the number of unique values grows, which requires more and more storage, and the newly observed values, which are often the most interesting, are given

no more weight than the old values. Gibbons and Mattias (1998) consider fixed-buffer-length sampling schemes that replace an element in a buffer (the current fixed-length sample) with a value that is not in the sample with a probability that depends on the buffer counts so far. The basic idea is that a newly observed value that recurs, and so is "hot," will eventually be included in the sample. Variants of this scheme use exponentially weighted moving averages of the buffer probabilities and include a buffer element for "other," which tracks the probability that the buffer does not include a recent item.

In choice-based and case-based sampling, whether for fixed populations or data streams, the probability of selection depends on an outcome like disease status. Case-based sampling is necessary when an outcome is so rare that simple random sampling would likely produce too few positive cases and an overwhelming number of negative cases. It is used in applications as diverse as medicine, ecology, and agriculture. For example, Suter et al. (2007) matched grassland plots with the poisonous grassland weed *Senecio jacobaea* and neighboring plots without *S. jacobaea* to determine environmental conditions conducive to the growth of *S. jacobaea*.

Event-based sampling, in which data are collected only when a signal exceeds a threshold or when an alarm sounds, is yet another kind of adaptive sampling. It is commonly used for analyses from engineering and Earth and planetary sciences for which storing all the data would be too costly or impractical. The threshold used depends on scientific knowledge about the size of interesting events. For example, the data rate on detectors in high-energy physics is so extreme that thresholds in hardware only allow 1 in 10 million events to pass to the next stage of the data system. The surviving data are then hierarchically sampled and filtered for further processing.

Size-biased sampling is similar: the probability of selection depends, either intentionally or not, on a measure of size, such as the duration of hospital stay or revenues of a corporation, that is related to the outcome. Size bias that is intentional can be removed during estimation and model fitting by weighting the data; unintentional size bias requires more specialized statistical analyses to remove (e.g., Vardi, 1982).

Specialized sampling techniques have evolved in ecology and evolutionary and environmental biology, and some of these are applied in large-scale applications. For example, estimation of wildlife population sizes has long used capture-recapture. The key idea is to capture animals, tag them, and then recapture the same population. The fraction of previously unseen animals in the recapture provides information about the population size. Capture-recapture designs can be quite elaborate and are now widely used in epidemiology.

There is also a growing literature on random sampling for simple networks and graphs; for example, Handcock and Gile (2010). The network is

represented as a binary matrix where element (i, j) is 1 if that pair of nodes is sampled, and 0 if not. Random sampling then chooses one such binary matrix randomly. For example, a node i may be chosen by simple random sampling and then all the nodes j that interact with i are included in the sample. Such sampling preserves some relationships, but it destroys many others. Sampling on networks is still in its infancy, and there are obvious opportunities for stratification, clustering, and adaptation (such as link tracing), with different consequences for estimation.

There are many more principled ways to sample data randomly. In theory, random sampling is easy, but in practice it is fraught with unforeseen complications, often due to a poor match between the sampling design and the structure and noise of the data.

Sampling according to known probabilities, whether equal or not, or whether all-at-once or sequentially, is called random sampling. Stratified sampling, case-based sampling, adaptive sampling, and length-biased sampling are all examples of random sampling. They all share the premise that the probability model that defines the sampling plan can be "unwound" to estimate a parameter of interest for the population at large, even if some observations, by design, were more likely to be sampled than others. Unequal weights give intentional sampling bias, but the bias can be removed by using the sampling weights in estimation or by estimating with regression or other models, and the bias allows attention to be focused on parts of the population that may be more difficult to measure. Finally, whether simple or not, random sampling leads to valid estimates of the reliability of the estimate itself (its uncertainty).

It is important to note that the analysis goals drive the sampling. Consider for example a repository of Web usage data. Sampling of search queries—probably by country, language, day of week, and so on—may make sense if the goal is to measure the quality of search results. On the other hand, if the goal is to measure user happiness, then one should sample across users, and perhaps queries within users. Or, it might be appropriate to sample user sessions, which are defined by nearly uninterrupted periods of activity for a user. At the other extreme, the sampling unit might be low-level timing events within queries. The decision about how to sample should follow the decision about what to sample.

Non-Random Sampling

Unfortunately, random sampling is not always practical. Members of stigmatized populations, like intravenous drug users, may hide the fact that they are part of that population, so the sampling frame is unknown. Some populations, like that of people with particular expertise or perspectives, may be difficult to identify. Snowball sampling, proposed by Goodman (1961), starts with a set of seeds that are known to belong to the hidden group. The seeds are asked to identify additional members of the group, then the newly identified members are asked to identify other group members, and so on. This process produces a sample without knowing the sampling frame. Goodman showed that if the initial seeds are random, then it is possible under some conditions to estimate relationships between people from snowball samples. Of course, there is a danger that the population is not well mixed, so only a non-representative sample can be reached from the initial seed. In contrast, random sampling is unbiased in the sense that two people in the same stratum, for example, have the same chance of being in the sample, even if their characteristics beyond those used for stratification are very different.

Heckathorn (1997, 2002) improved snowball sampling by adding more structure to sample recruitment. Each recruit at each stage is given a set number of coupons, a constraint that controls the rate of recruitment in sub-groups (as in stratification), reduces the risk of sampling only a dominant subgroup, and makes it easier to track who recruited whom and how large a social circle each recruit has. This respondent-driven sampling (RDS) has largely replaced snowball sampling. RDS is not random sampling because the initial seeds are not random, but some have suggested that the mixing that occurs when there are many rounds of recruitment or when the recruits added at later stages are restricted to those that have been given exactly one coupon induces a pseudo-random sample. However, difficulty in recruiting initial seeds may still lead to bias, and there is controversy about the reliability of estimates based on RDS even after correcting for the size of the networks of the initial seeds (e.g., Goel and Salganik, 2010). Finally, note that both RDS and snowball sampling often focus on estimation for an outcome in a population that is reached through a social graph rather than estimation of properties of the graph itself.

Sparse Signal Recovery

There are alternatives to sampling when collecting all the data is impractical, especially in applications in the physical sciences and engineering. In sparse signal recovery, a data vector x is multiplied by a matrix M, and only the resulting $y = Mx$ is retained. If x has only a few large components, then the matrix M can have considerably fewer rows than columns, so the vector y of observations is much smaller than the original data set without any loss of information. This kind of data reduction and signal recovery is known as compressed sensing, and it arises in many applications, including analog-to-digital conversion, medical imaging, and hyperspectral imaging, with a variety of different features that depend on the application (Candès and Tao, 2005; Donoho, 2006).

In signal and image processing, sparse recovery is done physically. For example, analog-to-digital converters sample analog signals, quantize the samples, and then produce a bit-stream. Hyperspectral cameras use optical devices to acquire samples that are then digitized and processed off-line. Streaming algorithms employ software-based measurement schemes.

Group testing is a special case of sparse signal recovery. Here the observations are taken on groups or pools of respondents, and only the combined response for the group is observed. Dorfman (1943), for example, combined blood samples from groups of army recruits to detect if any of them had syphilis. Because syphilis was uncommon, most groups would be negative, and no further testing was needed.

Sampling for Testing Rather Than Estimation

The discussion so far has focused on sampling for estimation, but it is also common to sample in order to test whether a set of factors and their interactions affect an outcome of interest. The factors under test are called *treatments*, and the possible values of a treatment are called *levels*. For example, two treatments might be "image used in ad campaign" and "webpage where the ad is shown." The levels of the first treatment might be the various images shown to the user, including the null value of "no ad," which is called the control. Often only a few levels of each treatment are tested. The decision about which test units get which combination of levels of the treatments is known as experiment design. Designed experiments go

back at least as far as the 1700s, but Fisher's book *The Design of Experiments* (1935) was the first to lay out principles for statistical experimentation.

The basic premise of experiment design is that it is much more efficient and informative to apply several treatments to the same unit under test, thus testing the effect of several treatments simultaneously, than it is to apply only one experimental factor to each test unit. That is, instead of testing whether the different levels of treatment A give different results on average, and then testing whether the different levels of treatment B give different results on average, both A and B are tested simultaneously by assigning every test unit a level of A and a level of B.

There are many kinds of experiment designs, just as there are many sampling designs. Perhaps the most common are fractional factorials that allocate test units to combinations of levels of treatment factors when the number of treatment levels plus the number of treatment interactions of interest exceed the number of units available for testing. In that case, various criteria are optimized to choose which combinations of treatment levels to test and how many units to test at each combination. Classical combinatorial theory is often used to find the assignment of treatments to experimental units that minimize the expected squared error loss (e.g., Sloane and Hardin, 1993).

Test units are often a simple random sample from the population of interest, and they are often randomly assigned levels of the factors to be tested. However, test units can be blocked (stratified) to control for confounding factors that are not under test, such as age. For example, test units might be stratified by country when testing the effects of different ad images and websites. Then comparisons are made within blocks and combined over blocks. Sequential testing and adaptive testing (e.g., multi-armed bandits) are also common in some areas, such as in medicine. These designs change the fraction of test units assigned to each combination of treatment factors over time, eliminating treatment levels as soon as they are unlikely to ever show a statistically significant effect. These ideas also extend to online experiments, where tests of different treatments are started and stopped at different times. That is, treatments A, B, and C may be tested on day 1 and treatments A, B, and D tested on day 2. If standard

design principles are followed, then the effects of the different treatments and their interactions can be teased apart (Tang et al., 2010).

Random sampling can be used not only to assign units to factors but also to evaluate which, if any, of the treatment factors and their interactions are statistically significant. As a simple example, suppose there is only one factor with two levels A and B, m units are assigned to A and n to B, and the goal is to decide if the mean outcome under A is different from the mean outcome under B after observing the difference D_{obs} of sample means. If the factor has no effect, so that there is no difference in A and B, then the A and B labels are meaningless. That is, D_{obs} should look like the difference computed for any random re-labeling of the units as m As and n Bs. The statistical test thus compares D_{obs} to the difference of sample means for random divisions of the observed data into A and B groups. If D_{obs} is larger or smaller than all but a small fraction of the mean differences obtained by random re-labeling, then there is statistical evidence that A and B have different effects on the mean outcome. On the other hand, if D_{obs} is not in the tail of the re-labeled differences, then one cannot be sure that the difference between A and B is meaningful. Such tests are called randomization or permutation tests. Tests that re-label a set of $m+n$ test units drawn with replacement from the original data are called bootstrapped tests; these behave similarly. Randomization and bootstrapped tests are nearly optimal in finite samples under weak conditions, even if the space of possible re-labelings is only sampled. In other words, valid tests can be constructed by approximating the sampling distribution of the test statistic with random re-sampling of the data.

Finally, it is not always possible to assign units to treatments randomly. Assigning students to a group that uses drugs to measure the effect of drug use on grades would be unethical. Or people may be assigned randomly to a treatment group but then fail to use the treatment. For example, a non-random subset of advertisers may be offered a new tool that should make it easier to optimize online ad campaigns, and a non-random subset of those may choose to use it. Or the treatment group may be identified after the fact by mining a database. Even though the treatment group is not random, valid tests can often be based on a random set of controls or comparison group, perhaps stratifying or matching so that controls and

"treated" were similar before treatment. Testing without randomization or with imperfect randomization falls under the rubric of observational studies.

CHALLENGES WHEN SAMPLING FROM MASSIVE DATA

Impressive amounts of data can now be collected and processed. But some tasks, like data visualization, still require sampling to tame the scale of the data, and some applications require sampling and testing methods that are beyond the state of the art. Following are a few examples of the latter.

Data from Participatory Sensing, or "Citizen Science"

There are billions of cell phones in the world. In participatory sensing, volunteers with cell phones collect location-tagged data either actively (taking images of overflowing trash cans, for example) or passively (reporting levels of background noise, pollutants, or health measurements like pulse rates). In principle, simple random sampling can be used to recruit participants and to choose which active devices to collect data from at any time and location, but, in practice, sampling has to be more complex and is not well-understood. Crowdsourcing, such as the use of micro-task markets like Amazon's Mechanical Turk, raise similar sampling questions (see Chapter 9). Challenges, which result from the non-stationary nature of the problem, include the following:

- *Sample recruitment is ongoing.* The goal is to recruit (i.e., sample mobile devices or raters for micro-task markets) in areas with fast-changing signals (hot spots) or high noise or that are currently underrepresented to improve the quality of sample estimates. Those areas and tasks have to be identified on the fly using information from the current participants. Recruiting based on information from other participants introduces dependence, which makes both sampling implementation and estimation more difficult. There are analogies with respondent-driven sampling, but the time scales are much compressed.

- *Some incoming samples are best omitted.* This is, in a sense, the inverse of the sample recruitment challenge just mentioned: how to decide when to exclude a participant or device from the study, perhaps because it produces only redundant, noisy, spotty, or perverse data.
- *Participants will enter and leave the sensing program at different times.* This hints of panel sampling with informative drop-out, but with much less structure. Note that the problem is non-stationary when the goals change over time or the user groups that are interested in a fixed goal change over time.
- *Data need not be taken from all available devices all the time.* Randomly sampling available devices is one option, but that could be unwieldy if it requires augmenting information from one device with information from others. Here, the problems are intensified by the heterogeneous nature of the data sources. Some devices may be much more reliable than others. Moreover, some devices may give data in much different formats than others.
- *Sampling for statistical efficiency and scheduling for device and network efficiency have an interplay.* How should sampling proceed to minimize communication (e.g., battery) costs? There may be analogies with adaptive testing or experiment design.
- *Data obtained from participatory sensing and crowdsourcing is likely to be biased.* The mix of participants probably does not at all resemble a simple random sample. Those who are easiest to recruit may have strong opinions about what the data should show, and so provide biased information. Samples may need to be routinely re-weighted, again on the fly, with the weights depending on the purpose of the analysis. New ways to describe departures from random sampling may need to be developed. Past work on observational studies may be relevant here.

For recent references on sampling and sampling bias in crowdsourcing or citizen science, see, for example, Dekel and Shamir (2009), Raykar et al. (2010), and Wauthier and Jordan (2011).

Data from Social Networks and Graphs

Statistically principled sampling on massive graphs, whether static or dynamic, is still in its infancy. Sampling—no matter how fine its resolution

—on graphs can never preserve all graph structure because sampling never preserves everything. But having some insight into some properties of a graph and its communities, and being able to characterize how reliable that insight is, may be better than having none at all. If so, then a set of

design principles for sampling from graphs for different kinds of analyses is needed.

Variants of respondent-driven sampling have been applied to graphs by randomly sampling a seed set of nodes and then randomly sampling some of the nodes they are connected to, with the selection depending on a window of time for dynamic graphs. Sometimes new random seeds are added throughout the sampling. Other methods are based on random walks on graphs, starting from random nodes, perhaps adding new seeds at random stages. With either approach, there are many open questions about tailoring sampling methods to the analysis task. Just as with smaller problems, there is likely to be no one right way to sample, but rather a set of principles that guide sampling design. However, these principles are not yet in place.

Entirely new ways to sample networks may be needed to obtain dense-enough subgraphs to give robust estimates of graph relationships. Traditional network statistics are highly sensitive to missing data or broken relations (e.g., Borgatti et al., 2006). For example, a rank order of nodes can be radically different if as little as 10 percent of the links are missing. New ways to sample that preserve node rankings are needed, as is theoretical understanding of the biases inherent in a strategy for network sampling. In particular, sampling designs that account for the network topology are needed. There has been some initial work (e.g., Guatam et al., 2008), but this topic is still in its infancy.

Finally, the difficulties in sampling networks are compounded when the data are obtained by crowdsourcing or massive online gaming. They are further intensified when individuals belong to multiple social networks, maintained at different sites. Answering a question as seemingly simple as how many contacts does an individual have is fraught with technical difficulty.

Experiment design for social networks is even less explored. Here the unit under test is not a node but a connected set of nodes. Choosing

equivalent sets of connected nodes that can be randomly assigned to the different levels of the treatments has only recently received attention (e.g., Backstrom and Kleinberg, 2011.) There is also a need for methods that allow principled observational studies on graphs. Aral et al. (2009) provide an early example that uses propensity scoring to find appropriate random controls when the treatment cannot be properly randomized.

Data from the Physical Sciences

The emergence of very large mosaic cameras in astronomy has created large surveys of sky images that contain 500 million objects today and soon will have billions. Most astronomers today use a target list from a large, existing imaging survey to select their favorite objects for follow-up observations. Because these follow-ups are quite expensive in terms of telescope time, various (mostly ad hoc) sampling strategies are applied when, for example, selecting galaxies for spectroscopic observations. Although it is relatively straightforward to create a sampling scheme for a single criterion, it is rare that a sample created for one purpose will not be later reused in another context. The interplay between often conflicting sampling criteria is poorly understood, and it is not based on a solid statistical foundation.

As a concrete example, both Pan-STARRS and the Large Synoptic Survey Telescope will provide deep multicolor photometry in co-added (averaged) images for several billion galaxies over a large fraction of the sky. This data set will provide an excellent basis for several analyses of large-scale structure and of dark energy. However, unlike the Sloan Digital Sky Survey (SDSS), these surveys will not have a spectroscopic counterpart of comparable depth. The best way to study structure, then, is to split the galaxy sample into radial shells using photometric redshifits—i.e., using the multicolor images as a low-resolution spectrograph. The most accurate such techniques today require a substantial "training set" with spectroscopic redshifits and a well-defined sample selection. No such sample exists today, nor is it clear how to create one. It is clear that its creation will require major resources (approximately 50,000-100,000 deep redshifits) on 8- to 10-m class telescopes.

Given the cost of such an endeavor, extreme efficiency is needed. Special care must be taken in the selection of this training set, because

galaxies occupy a large volume of color space, with large density contrasts. Either one restricts galaxy selection to a small, special part of the galaxy sample (such as luminous red galaxies from the SDSS), or one will be faced with substantial systematic biases. A random subsample will lack objects in the rare parts of color space, while an even selection of color space will under-sample the most typical parts of color space. The optimum is obviously in between. A carefully designed stratified sample, combined with state-of-the-art photometric redshift estimation, will enable many high-precision cosmo-logical projects of fundamental importance. This has not been attempted at such scale, and a fundamentally new approach is required.

Time-domain surveys are just starting, with projected detection cardinalities in the trillions. Determining an optimal spatial and temporal sampling of the sky (the "cadence") is very important when one operates a $500 million facility. Yet, the principles behind today's sampling strategies are based on ad hoc, heuristic criteria, and developing more optimal algorithms, based on state-of-the-art statistical techniques, could have a huge potential impact.

As scientists query large databases, many of the questions they ask are about computing a statistical aggregate and its uncertainty. Running an SQL statement or a MapReduce task provides the "perfect" answer, the one that is based on including all the data. However, as data set sizes increase, even linear data scans (the best case, because sometimes algorithms have a higher complexity) become prohibitively expensive. As each data point has its own errors, and statistical errors are often small compared to the known and unknown systematic uncertainties, using the whole data set to decrease statistical errors makes no sense. Applying an appropriate sampling scheme (even incrementally) to estimate the quantities required would substantially speed up the response, without a loss of statistical accuracy. Of course, scientific data sets often have skewed distributions; not everything is Gaussian or Poisson. In those cases, one needs to be careful how the sampling is performed.

REFERENCES

Aral, S., L. Muchnik, A. Sundararajan. 2009. Distinguishing influence based contagion from homophily driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences U.S.A.* 106:21544-1549.

Backstrom, L., and J. Kleinberg. 2011. Network bucket testing. Pp. 615-624 in *Proceedings of the 20th International World Wide Web Conference*. Association for Computing Machinery, New York, N.Y.

Borgatti, S., K.M. Carley, and D. Krackhardt. 2006. Robustness of centrality measures under conditions of imperfect data. *Social Networks* 28:124-136.

Candès, E., and T. Tao. 2005. Decoding by linear programming. *IEEE Transactions on Information Theory* 51:4203-4215.

Dekel, O., and O. Shamir. 2009. Vox Populi: Collecting high-quality labels from a crowd. Pp. 377-386 in *Proceedings of the 22nd Annual Conference on Learning Theory (COLT)*. Available at http://www.cs.mcgill.ca/~colt2009/proceedings.html.

Donoho, D.L. 2006. Compressed sensing. *IEEE Transactions on Information Theory* 52: 1289-1306.

Dorfman, R. 1943. The detection of defective members of large populations. *The Annals of Mathematical Statistics* 14:436-440.

Fisher, R.A. 1935. *The Design of Experiments.* Macmillan, New York, N.Y.

Fisher, R.A., and F. Yates. 1938. *Statistical Tables for Biological, Agricultural and Medical Research*, 3rd ed. Oliver and Boyd, London. Pp. 26-27.

Gautam, D., N. Koudas, M. Papagelis, and S. Puttaswamy. 2008. Efficient sampling of information in social networks. Pp. 67-74 in *Proceeding of the 2008 ACM Workshop on Search in Social Media (SSM '08)*. Association for Computing Machinery, New York, N.Y.

Gibbons, P.B., and Y. Matias. 1998. New sampling-based summary statistics for improving approximate query answers. Pp. 331-342 in *Proceedings of the 1998 ACM International Conference on Management of Data (SIGMOD)*. Association for Computing Machinery, New York, N.Y.

Goel, S., and M.J. Salgonik. 2010. Assessing respondent-driven sampling. *Proceedings of the National Academy of Sciences U.S.A.* 107:6743-6747.

Goodman, L.A. 1961. Snowball sampling. *The Annals of Mathematical Statistics* 32:148-170.

Handcock, M.S., and Gile, K. 2010. Modeling social networks from sampled data. *The Annals of Applied Statistics* 4:5-25.

Heckathorn, D. 1997. Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems* 44:4174-199.

Heckathorn, D. 2002. Deriving valid population estimates from chain-referral samples of hidden populations. *Social Problems* 49:11-34.

Raykar, V.C., S. Yu, L.H. Zhao, G.H. Valadez, C. Florin, L. Bogoni, and L. Moy. 2010. Learning from crowds. *Journal of Machine Learning Research* 11:1297-1322.

Sloane, N.J.A., and R.H. Hardin. 1993. *Journal of Statistical Planning and Inference* 37: 339-369.

Suter, M., S. Siegrist-Maag, J. Connolly, and A. Lüscher. 2007. Can the occurrence of *Senecio jacobaea* be influenced by management practice? *Weed Research* 47:262-269.

Tang, D., A. Agarwal, D. O'Brien, and M. Meyer. 2010. Overlapping experiment infrastructure: More, better, faster experimentation. Pp. 17-26 in *Proceedings of the 16th Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, New York, N.Y.

Vardi, Y. 1982. Nonparametric estimation in the presence of length bias. *The Annals of Statistics* 10:616-620.

Vitter, J.S. 1985. Random sampling with a reservoir. *ACM Transactions on Mathematical Software* 11:37-57.

Wauthier, F.L., and M.I. Jordan. 2011. Bayesian bias mitigation for crowdsourcing. Pp. 1800-1808 in *Proceedings of the Conference on Neural Information Processing System, Number 24*. Available at http://machinelearning.wustl.edu/mlpapers/papers/NIPS2011_1021.

The National Academies of Sciences, Engineering, and Medicine
500 Fifth St., NW | Washington, DC 20001