Contents

1

Statistical Inference and Differential Privacy

Jordan Awan

Purdue University

Ruobin Gong

Rutgers University

CONTENTS

1.1 Introduction: statistical inference in the presence of privacy requirements

The primary objective of statistical inference is to learn about certain aspects of a population. This population may either be real or hypothetical, as well as finite or infinite. Often it is the case that the inference is performed on a smaller sample – a database – that is reasonably representative of the population. The analyst identifies the unknown quantity of interest, posits a statistical model that dictates the probabilistic relationship between the quantity of interest and the observable data, and lets the observed data guide the inferential conclusions. For the statistical inference to be reliable and trustworthy, it must be performed in a manner that appropriately accounts for the various sources of uncertainty in the data collection, as well as any variability due to data processing, including (but not limited to) the need for privacy protection.

Because the privacy mechanism injects additional randomness into the data products, analyzing them as if they had not gone through privacy protection may lead to erroneous statistical conclusions. In this chapter, we discuss how to perform statistical inference based on privacy-protected data products. Our focus is on the methodology for carrying out statistical inference based on privatized data to the best extent possible, where "best" speaks to the quality of the inferential conclusions. Key to the possibility of good inference is when privacy protection can be performed in a *transparent* manner. Knowledge about the privacy mechanism facilitates informed and principled downstream statistical analysis.

Crudely speaking, the more privacy we wish to preserve, the less information we should be able to learn. That is the intuition behind the *privacy-utility trade-off* [2], the idea that the protection of the information of individuals belonging to a population negatively impacts the ability to draw accurate statistical conclusions about that population.

The privacy-utility trade-off, pitting the learning of population-level features against the regulation of individual-level uncertainties, is as old as the subject of statistical inference itself. Statistical inference is the science of taming noise and variability at higher resolutions, while revealing meaningful signal and regularity at lower resolutions. Therefore, while the differential privacy constraint protects from learning about the individuals in the dataset, it need not prevent the analyst from learning population-level characteristics to a satisfactory level of accuracy.

In order to have a systematic discussion about balancing this trade-off, we need to endorse an analytical conceptualization of both privacy and statistical utility. How much privacy might we gain, if we can tolerate a fixed level of uncertainty in the conclusions? On the other hand, if the statistical conclusion must meet a certain quality level, what implications does it have on the privacy of the individuals?

The definition of differential privacy is conducive to the quantified balance between the needs for privacy protection and statistical utility. As discussed in Chapter [formal-DP], differential privacy provides a notion of plausible deniability for individual participants of the database. It ensures that the analyst or attacker cannot know, to a high degree of certainty, about what data any single participant contributed, or whether they contributed anything at all. Within a given data curation setting, differential privacy quantifies the extent of protection through the privacy loss budget. It is a sliding scale that allows the data curator to choose the degree of privacy protection.

The concept of statistical utility, on the other hand, can have a varied interpretation. Its meanings would differ depending on the specific goals of the inference task. We begin the chapter with a discussion of the various important and practical notions of statistical utility in Section 1.1.1. As we will see in Section 1.1.2, the differential privacy requirement can be understood as a constraint on the statistical information permissible in the data product through the lenses of likelihood, frequentist, and Bayesian inference, leading to natural ways of understanding the impact of privacy on statistical utility within these inference schemes.

The remainder of the chapter is organized as follows. Section 1.2 sketches the methodological elements for drawing statistical inference from differentially private information. We motivate a general marginal likelihood construction by viewing the confidential data as latent observations, and underscore the importance of a *transparent* privacy mechanism – that is, the probabilistic characteristic of the privatized statistics relative to the confidential data being known to the analyst. Section 1.3 provides an overview of the existing methodologies for differentially private statistical inference. We make a practical distinction between methodologies designed for the *dissemination*-based versus *query*-based data access modes, and remark that the former setting is more general. Section 1.4 presents several open challenges in drawing statistical inference under differential privacy, including disclosure risk control and individual feature prediction under differential privacy, the presence of invariants and logical constraints to privatized data products, the need for clamping queries with unbounded sensitivity, as well as considerations around identifiability and estimability of parameters under privacy constraints, and consistency of estimators that result from them. Section 1.5 concludes with a discussion.

1.1.1 Statistical utility

Statisticians rarely speak of the utility of a data product as a standalone concept. The discussion is only stipulated as we seek for the balance between good privacy protection and high-quality statistical inference. Therefore, how to frame statistical utility as a measurable quantity comes down to the question: what good qualities do we value the most in statistical inference?

An intuitive notion of statistical utility measures how close the private outputs are to the non-private outputs for the database at hand. These measures are commonly employed in the computer science literature and are applied to descriptive statistics. For example, the evaluation of the Disclosure Avoidance System (DAS) of the 2020 Decennial Census P.L. 94-171 redistricting data product includes the *mean absolute error* (MAE) and the *mean absolute percent error* (MAPE) as accuracy measures for the population counts at various geographic levels [73]. These measures are straightforwardly defined in terms of the values of the original and the privatized statistics. They are useful measures of the extent of perturbation in the published data product incurred by the privacy mechanism.

However, simple closeness measures may not reflect how well the private output can be used to learn about population-level quantities, as is often the goal in statistical inference. A second class of statistical utility definitions views the private statistic as an estimator of a population parameter of interest. In this view, we can ask whether the statistic is *unbiased*, that is whether the error due to privacy is zero in expectation, and whether it is *consistent* for the parameter of interest, that is whether the error due to privacy would become asymptotically negligible if more and more data could be collected. For inference, it is important that we measure the *variance* of the estimator, and more informatively its *mean squared error* (MSE). The MSE of an estimator is equal to the sum of its squared bias and its variance. A smaller MSE is typically a sign of a high-quality estimator, for it gives the assurance that on average, the amount that the estimator deviates from its estimand, as measured in squared distance, is small. We may further inquire whether the estimator is optimal in some sense. One such sense of optimality is *minimax*, that is, whether the estimator poses the smallest maximal risk among all alternative estimators. Along these lines, there have been several promising results [71, 20, 29], which established that a wide class of population-level parameters can be estimated, under differential privacy protection, with asymptotically negligible noise.

Beyond simply estimating an unknown quantity, we may wish to conduct uncertainty quantification about the population parameter of interest based on the private statistic. This could take the form of a hypothesis test or a confidence region in the frequentist tradition, or a posterior distribution in the Bayesian tradition. We will typically require these inferences to be externally valid, in the sense of meeting the nominal Type I error and coverage rates, or being calibrated to the hypothetical replications of the data collection process. This is perhaps the most challenging goal of statistical inference subject to differential privacy. Later in this chapter, we will see that performing valid statistical inference under this setting requires an understanding of the marginal likelihood of the parameter given the privatized statistic.

A common and important task for statistical modeling is to perform individualized prediction. Statistical utility in this context is measured in terms of the accuracy of the predictions about each individual's feature as generated by the model, assessed against the actual feature that the individual possesses. The tension between privacy and utility is particularly acute in individualized prediction, because both are formulated, hence measured, directly in terms of the individual's feature. We remark, however, that even if the predictive model is trained subject to the differential privacy guarantee, it is still possible to make accurate predictions about the individual's class or response. Typical situations in which this may arise are when the model can access other public information (such as useful predictors of the individual) or common expert knowledge. For example, if a scientific study shows a reliable connection between smoking and cancer, and it is publicly known that a person named Bob is a chain smoker, one can infer from the study that Bob may develop cancer with high probability. Note that this inference about Bob does not constitute a violation of the differential privacy guarantee, even if it implies a high *disclosure risk* for Bob's cancer status. Indeed, this inference can be reached even if Bob is not part of the scientific study. We discuss the issue of privacy and individualized prediction further in Section 1.4.

Statistical inference differentiates itself from other modes of learning by a strong emphasis on the quantification of uncertainty. Statistical methodologies are concerned with not only taming uncertainty as much as possible, but also with truthfully measuring uncertainty in any given situation. In other words, we try hard to learn, and at the same time stay honest about how much we can and cannot learn. Ultimately, the utility of statistical inference is reflected in the positive impact by the evidence-based decisions made in the downstream. Any sacrifice that is made to statistical utility in lieu of privacy must also be quantified within the same context.

1.1.2 The statistical meaning of differential privacy

Loosely speaking, the differential privacy requirement says that for two neighboring databases that differ by one individual's contribution, the probability distributions of the privatized outputs are similar. This requirement poses a constraint on the statistical information that the data product may contain. This section offers several statistical interpretations of differential privacy, associating it with hypothesis testing, likelihood, and Bayesian inference. These views help understand the impact of differential privacy under the most commonly employed statistical inference schemes.

The first and most straightforward interpretation of differential privacy is that it poses a bound on the divergence between the two output distributions relating to the two underlying neighboring databases. This perspective is most commonly referenced in the computer science community, and offers a somewhat intuitive notion of privacy. If the output distributions are similar, as measured in terms of some divergence, then it must be difficult to tell which distribution the realized sample actually came from. This viewpoint has led to alternative notions of differential privacy, including Rényi differential privacy, concentrated differential privacy, and truncated concentrated differential privacy [34, 19, 64].

That the two output distributions are difficult to tell apart can be formulated with the language of statistical hypothesis testing. Specifically, differential privacy amounts to requiring that, given the privatized output, any hypothesis test attempting to discern which of the two databases with which we began has bounded power as a function of the Type I error. [80] first noted that differential privacy was equivalent to imposing a bound on the power of any hypothesis test of whether an individual participated in the database or not, based on the privatized output, and this perspective was further explored by [51]. Recently, [26] use this viewpoint to propose a family of differential privacy guarantees parameterized by the different choices of bounds on the Type I and Type II errors. This hypothesis testing viewpoint shows what an adversary is able to accomplish based on the privatized output, and formalizes the "plausible deniability" afforded to the participating individuals. That is, it is difficult to tell whether any of them participated in the dataset at all.

The classic ϵ-differential privacy can also be interpreted as imposing a bound on the likelihood ratio between the distributions of outputs when using two databases differing in one individual. This view has implications on both the testing and the estimation perspectives of statistical inference. As the likelihood ratio statistic is known to be the most powerful test for discerning two simple hypotheses, this interpretation highlights the link between differential privacy and hypothesis testing. On the other hand, the bound on the likelihood ratio also indicates that an analyst's likelihood function remains *similar* when one person's data is changed in the original database, where the similarity is up to a factor of $\exp(\epsilon)$. This means that ϵ-differential privacy poses a bound (by $\pm \exp(\epsilon)$ fold, to be precise) on the *Bayes factor*, that is the ratio between the posterior odds and the prior odds, harbored by the analyst regarding whether an individual is included or excluded from the actual database. Importantly, the bound holds regardless of the prior information that the analyst may have, again limiting the extent to which anyone can learn about the participation of the individual.

Note that differential privacy can only provide a relative guarantee to an individual's privacy, and does not serve as an absolute guarantee. The differential privacy guarantee is only relative, in the sense that whether an individual is present in the database or not, the resulting outcome does not change much. However, as discussed in Section 1.1.1, it is still possible to make accurate inference about the status of an individual based on the privatized output. Specifically, the smoking and cancer example discussed there would stand even if the individual in question is not present in the database at all. From a Bayesian perspective, while differential privacy bounds the ratio of the likelihood function whenever one individual's data is changed, it does not offer an absolute bound on the posterior distribution. [12] offers an extended analysis on the bounds that ϵ-differential privacy imposes on the appropriate likelihood function and other key quantities in frequentist and Bayesian inferences. Under mild conditions, these bounds are generally applicable to arbitrary parameters, data generating models and priors, and are non-vacuous in finite sample settings, shedding light on the limit of statistical learning from ϵ-differentially private data products.

Finally, we remark that there is a deep connection between differential privacy and robust statistics. Since differential privacy regulates the privacy mechanism in such a way that it not depend too strongly on any one individual's data, it is related in essence to statistical robustness notions such as the *breakdown point* and the *influence function*. [33] and [5] pointed out that robust estimators can be easily modified to satisfy differential privacy. The connection to robustness also allows for differential privacy to enable sequential data analysis via the *reusable holdout* [32], to avoid overfitting to the data and to enhance research reproducibility.

1.2 Inference through the lens of formal privacy

In this section, we discuss the methodological elements of drawing statistical inference from differentially private information. As we will see, good statistical inference from a privatized output hinges crucially on the knowledge about the probabilistic specification of its generation process. Therefore, the transparent specification of the privacy mechanism is a prerequisite to reliable inference, a point that we underscore at the outset.

1.2.1 The importance of transparent privacy

A great strength of differential privacy is that it is capable of providing security without obscurity. The privacy guarantee does not require the privacy mechanism itself to be secret, and the mechanism can be safely published along with the privatized output, with no additional threat to the confidentiality of individuals belonging to the database. In fact, the transparent specification of the mechanism is precisely what allows for the verification of its guarantee, a fundamental premise that allowed for the flourishing literature on differential privacy mechanism design, which caters to a variety of use cases while delivering better utility qualities. The reader is referred to Chapter [popular-DP-mechanism] of the handbook for more examples.

From the statistical point of view, we say that a privacy mechanism is *transparent*, if the conditional probability distribution of the privatized query z, given the underlying database x, is fully known to the data analyst who has access to z. We denote this conditional distribution as $\eta\left(z \mid x\right)$.

The transparency of the privacy mechanism plays a crucial role in conducting reliable statistical inference and uncertainty quantification based on its output [43]. The transparency property of differential privacy draws a stark contrast to the philosophy of traditional statistical disclosure limitation, where a privacy protection procedure (such as swapping or noise infusion) is typically regarded as a secret and is not publicly disclosed. If the privacy mechanism is not known, it poses various challenges to adequately analyzing the privatized output. For example, it may be the case that the analyst's intended model parameter – otherwise identifiable – becomes unidentifiable with privacy protection, yet the analyst would not know how to ascertain that. The best the analyst can do is to make assumptions, or guesses, about necessary aspects of the privacy mechanism. In the case of swapping, these may include the swap keys and swap rates. In the case of noise infusion, these may include the distribution, scale, and dependence structure for the noise variables. If these assumptions cannot be verified or supported by the curator's documentation, as often is the case with obscure privacy mechanisms, the statistical analysis that ensues may well be inaccurate and potentially invalid.

Because the specifications of a differential privacy mechanism can be made available to the data analyst, this information should and must be incorporated into the modeling procedure wherever possible. This would ensure the correct propagation of the additional uncertainty introduced for privacy, and the derivation of valid statistical inferences. For a well-specified differentially private mechanism, it is possible to write down a full generative model for the published privatized data conditional on the original confidential data. Further assuming a sampling model for the unobserved original data, any noise or perturbation due to privacy can be properly accounted for in the marginal likelihood for the parameters of interest given the privatized output. This allows for the appropriate accounting of errors, and is the foundation for valid statistical inference based on privatized data. In what follows, we describe in detail the analysis approach under transparent privacy.

1.2.2 Privacy mechanism and the marginal likelihood

In a statistical analysis task without privacy concerns, the analyst typically specifies a model for the data that depends on the unknown parameters, and possibly also any prior information they have. However, when a privacy mechanism is applied, the analyst no longer sees the original data, but only the privatized outputs. At this point, the original data becomes missing data, or latent variables in the analyst's model.

Let $f_X(\cdot \mid \theta)$ be the model for the private data X depending on the parameter θ. As introduced previously, let $\eta(\cdot \mid x)$ be the privacy mechanism through which the privatized data Z is probabilistically generated based on the original data $X = x$. Then the marginal model for Z is

$$p(z \mid \theta) = \int_{\mathscr{X}} \eta(z \mid x) f_X(x \mid \theta)\, dx. \tag{1.1}$$

This quantity is precisely the marginal likelihood of θ given only the privatized output Z, which was first identified in [81]. In the special case that the privacy mechanism adds an independent noise to a statistic $T(X)$, where the noise is drawn from the distribution $\eta(\cdot)$, then (1.1) takes the special form

$$p(z \mid \theta) = \int_{\mathscr{X}} \eta(T(x) - z) f_X(x \mid \theta)\, dx = \int \eta(t - z) f_{T(X)}(t \mid \theta)\, dt, \quad (1.2)$$

which we see is the convolution of the posited distribution of $T(X)$ given θ, and the noise distribution η. Many, but not all, differential privacy mechanisms are independent additive mechanisms, with noise taking distributions of Laplace, double geometric, Gaussian, t, and so on (e.g., [9, 68, 40]). On the other hand, there are also many mechanisms that do not have an additive structure, for which $\eta(z \mid x)$ remains in the more general form of (1.1). These include the classic randomized response mechanism [79], the exponential mechanism [63] and the objective perturbation [58].

The marginal likelihood $p(z \mid \theta)$ is the foundation to the statistical inference problem using privatized statistics. It incorporates both the sampling (or modeling) uncertainty about the parameter of interest θ, as reflected in f, and the uncertainty due to privacy, injected by the probabilistic mechanism η. The marginal likelihood describes the sampling properties of z given θ, guiding frequentist and likelihood estimation. Further combined with a prior distribution on θ, one can also derive the Bayesian posterior distribution of θ given z, that is

$$p(\theta \mid z) \propto \pi(\theta)\, p(z \mid \theta),$$

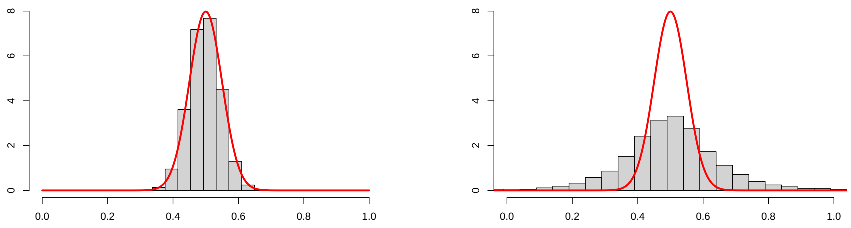
for the analyst's choice of prior $\pi(\theta)$.

Note that the representation of the marginal likelihood in (1.1) involves an integral over the entire space of possible input databases $x \in \mathscr{X}$. This could be a demanding task if no low-dimensional summary statistic or other types of tractable simplification exists. As Section 1.2.3 will discuss, inference strategies for privatized statistics are invariably designed around taming $p(z \mid \theta)$ in some way, be it analytically, computationally, or approximately.

1.2.3 Inference strategies for privatized statistics

In the previous section, we established that the likelihood function of a population parameter given a privatized statistic is the marginal likelihood function. Computing the marginal likelihood function requires integrating over the original dataset x, and can be potentially intractable when the dimension of the dataset is large. This often requires the use of either sampling techniques or approximation strategies, and sometimes a combination of the two.

Consider one of the simplest problems where each individual contributes a single binary value, modeled as a Bernoulli experiment. In this case, the marginal likelihood can be evaluated exactly. As the total number of "ones" is a sufficient statistic, [7] showed that evaluating the marginal likelihood only requires a single summation, which has running time that is linear in the



(a) Histogram of $\overline{x}$ over the 10000 replicates, along with the asymptotic distribution (1.3), in red.

(b) Histogram of $\overline{x}$ + Laplace $(1/(\epsilon n))$, with $\epsilon = .1$, along with the asymptotic distribution (1.4), in red.

FIGURE 1.1: Comparison of asymptotic approximating distributions in private and non-private settings. Data are drawn i.i.d. from Bern $(.5)$, with sample size $n = 100$, and 10000 replicates.

sample size. In the local model, [60] give conditions for the marginal likelihood to be tractable when using the Gaussian mechanism.

In cases where the marginal likelihood cannot be computed directly, other techniques are needed. It may be tempting to ignore the privacy mechanism, or to apply traditional statistical asymptotics to approximate the sampling distribution of a differentially private statistic. These approaches are not entirely unjustified. Because the noise introduced for privacy is often asymptotically negligible compared to the error due to sampling (e.g., [71, 20]), the asymptotic sampling distribution of many differentially private statistics is the same as their non-private counterpart, such as noted by [76, 36].

Unfortunately, while the noisy introduced for privacy is often asymptotically negligible, in finite sample sizes, it often results in inaccurate asymptotic approximations [76]. In Example 1, we present a case based on [76], which shows that standard asymptotic techniques can result in unacceptably poor approximations when applied in the differential privacy setting.

Example 1 (Classical asymptotics with DP)**.** *Let* $\mathbf{X} \in [0,1]^n$ *be a sequence of random variables drawn i.i.d. from a distribution* F *with known variance* σ^2*. The classical estimator of the mean* μ *of* F *is* $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n x_i$*, which by the central limit theorem has the asymptotic distribution*

$$\left(\sqrt{n}/\sigma\right)\left(\hat{\mu} - \mu\right) \xrightarrow{d} \mathcal{N}\left(0, 1\right). \tag{1.3}$$

We approximate the sampling distribution of $\hat{\mu}$ *as* $\mathcal{N}\left(\mu, \sigma^2/n\right)$*. To satisfy privacy, note that* $\hat{\mu}$ *has* sensitivity $1/n$*. Thus a simple mechanism to privatize* $\hat{\mu}$ *is to add independent noise distributed as* $N \sim$ Laplace $\left(\frac{1}{\epsilon n}\right)$*. Then the private output is* $\widetilde{\mu} = \hat{\mu}\left(\mathbf{X}\right) + N$*, which satisfies* ϵ*-DP. Note that since* $N =$

$O_p\left(1/n\right)$, we have the same asymptotic distribution for $\widetilde{\mu}$:

$$\left(\sqrt{n}/\sigma\right)\left(\widetilde{\mu}-\mu\right)\overset{d}{\to}\mathcal{N}\left(0,1\right). \tag{1.4}$$

Despite this asymptotic property, we see in Figure 1.1 that the normal approximation (red density) is considerably less accurate for the true sampling distribution of $\widetilde{\mu}$ in panel (b) than that of $\hat{\mu}$ in panel (a).

It is worth noting that in non-private settings, a sample size of $n = 100$ is usually large enough for the central limit theorem to give good approximations. While the privacy noise is asymptotically negligible, it is not negligible in finite samples. Example 1 emphasizes the need to explicitly incorporate the privacy noise into the statistical inference procedure.

While traditional asymptotics are often not accurate enough in the presence of a privacy mechanism, there have been several works that incorporate the privacy mechanism with traditional asymptotics, avoiding the problem of Example 1 (e.g., [37, 77, 38]). In particular, [76] developed an alternative asymptotic regime which, under certain conditions, produces approximations for differentially private statistics that are at least as accurate as their non-private counterparts.

A different class of approaches is based on the idea of producing samples from the marginal likelihood. Often the model for the private data is a generative model, meaning that given a parameter value, it is relatively easy to sample a new dataset. Furthermore, the privacy mechanism is fully specified, and is usually designed to be easily sampled. This structure of the privacy problem makes sampling from the marginal distribution of the privatized statistic relatively easy. By viewing the original data as latent variables, Markov chain Monte Carlo (MCMC) approaches can be used to sample from the marginal likelihood [81, 16, 17, 50]. In addition, [42] shows that Monte Carlo Expectation Maximization (MCEM) and approximate Bayesian computation (ABC) can be used to produce samples that are exact with respect to the marginal likelihood and the Bayesian posterior.

The ability to sample the differentially private statistic given a parameter can also be used for parametric bootstrap inference. [36] show that this approach gives accurate confidence intervals for a variety of problems, with higher accuracy than asymptotic techniques. The parametric bootstrap has also been used to approximate the sampling distribution in many private hypothesis testing problems [37, 3]. [6] develop an alternative method to approximately sample conditional on a differentially private statistic provided that it is asymptotically efficient. They show that this can allow one to test certain hypothesis based on a differentially private statistic. [11] showed that simulation-based inference can be used to produce confidence intervals and hypothesis tests with guaranteed coverage/type I error for general parametric models. Similar simulation-based inference methods can be used to de-bias a privacy mechanism [45]; see Section 1.4 for a discussion on biases in privacy mechanisms.

Another computational technique that has been successfully applied to the marginal likelihood is variational Bayesian analysis [52]. Variational techniques are an alternative MCMC methods, which work with a non-asymptotic approximation to the target distribution.

1.3 Methodologies for differentially private statistical inference: an overview

When tackling the problem of statistical inference under differential privacy, it is important to distinguish between the different modes of data access [49], due to their distinct implications on the downstream inference task. For the purpose of this chapter, two modes of data access are worth highlighting. Under the *query*-based mode, the data curator is interested in performing the statistical inference or in directly enabling it. In this case, the curator can carefully choose which private statistics to evaluate, and design these queries specifically for the statistical inference task at hand. For example, if we were the disclosure avoidance team at Google, Apple, or the Census Bureau, we would have the freedom to tailor our privacy mechanism to the statistical questions that we want to answer.

In contrast is the *dissemination*-based mode. Here, the data analyst is a separate entity from the data curator, who is entirely uninformed of the analyst's intent. The curator chooses what private statistics to release, without knowing the statistical questions the data analyst would like to pursue or the models they would like to fit. The curator may anticipate certain popular use cases, but cannot foresee all possible ways in which the private statistics will be analyzed. In this case, the analyst has no direct control over the privacy mechanism, and is left to perform statistical inference given the private output as provided by the curator. For example, if we were a team of social scientists interested in analyzing the Census Bureau's differentially private public data products, we would be situated in the dissemination mode.

Every technique developed under the dissemination mode is also applicable to the query mode. While the converse is not generally true, there have been several interesting works which showed that optimal differential private inference in the query mode often coincide with techniques from the dissemination mode.

We remark that the query vs. dissemination modes should not be confused with the central vs. local privacy models. The distinction between the central and the local models concerns the differential privacy guarantee and the nature of the privacy algorithms. On the other hand, the distinction between the query and the dissemination modes of data access concerns whether the analyst has the freedom to choose their own private queries, or are left to work with the private queries chosen and computed by someone else. It is possible to be in either the query or dissemination modes, regardless of whether the privacy model is central or local.

In addition, a database may offer more than one mode of access to its users. For example, the *tiered access* model would allow query-based access by certain approved data analysts and dissemination-based access by the general public. This hybrid approach is seen as a viable model of data access in many use cases within official statistical agencies [67, 47].

1.3.1 Inference under dissemination-based data access mode

The simpler of the two modes of data access is the dissemination mode. In this case, the analyst has no control over the privacy mechanism, and in that sense, their work is of a reduced degree of complexity. Here, the analyst is given a differentially private output, and is tasked with performing the statistical inference of interest based on it.

One approach that gives reliable statistical inference under the dissemination mode is the parametric bootstrap, which is also arguably the simplest blackbox approach. Regardless of the differentially private mechanism used, as long as the mechanism is fully specified, it allows the analyst to sample a new differentially private statistic given a parameter estimate. This is the basis for the parametric bootstrap inference. [36] show that this approach gives accurate confidence intervals for exponential families and regression models. A more complicated sampling approach to inference is simulation-based inference, which [11] showed can give confidence intervals and hypothesis tests with provable coverage.

When the privacy mechanism admits an explicit expression of the conditional distribution of the private query given the confidential one, [42] shows that exact samples from the posterior distribution can be obtained by a modified approximate Bayesian computation (ABC) scheme. This approach involves sampling from the data generating model, and accepting a sample when it gives a synthetic statistic that is close to the observed private statistic, where closeness is assessed using to the conditional distribution.

In a local model setting where noise is added directly to the values in the original dataset, measurement error models can be employed to perform correct inference on the parameters. Examples of methods that fit in this framework are randomized response and post randomization method (PRAM), which perturb the data on a record-by-record basis [44]. Measurement error models have been successfully used in this setting to perform Bayesian inference on regression [41] and small area estimation [66]. Similarly, the EM algorithm can be applied to obtain unbiased parameter estimates when the data was privatized by PRAM [82].

The marginal likelihood perspective as described in Section 1.2.2 is generically applicable in the dissemination mode. This is first noted in [81], where they use MCMC methods to sample from the marginal posterior distribution in a logistic regression. [52] apply variational methods to work with the posterior distribution in a naïve Bayes classifier.

While we discussed earlier that traditional asymptotic techniques do not necessarily deliver valid finite sample statistical inference in differentially private problems, there have been some works that successfully incorporate the privacy noise into the asymptotic approximation. [74] is one of the earliest papers on statistical inference for differential privacy. They show that in a dataset modeled as independent Bernoulli experiments, with Laplace noise added to the sum of the observations, the sampling distribution can be approximated either with the convolution of a Gaussian and a Laplace random variables, or with a normal approximation with an inflated variance. They use these approximations to derive sample size calculations based on the differentially private outputs, and to perform tests of the population proportion and tests of independence where Laplace noise is added to the cells of a contingency table. [37] develop alternative differentially private tests for independence and goodness-of-fit tests for contingency tables, when noisy counts are observed. They offer both asymptotic and Monte Carlo methods for calibrating the Type I errors. [77] discuss an asymptotic framework for approximating the sampling distribution of independence and goodness-of-fit test statistics based on noisy data, which are further developed in [76].

1.3.2 Inference under query-based data access mode

Under the query-based data access mode, the data analyst has the ability to design the differentially private statistical queries that will be evaluated on the confidential database. This allows the analyst to tailor the queries to the statistical task of interest.

Point estimation is a central statistical task, one that has received a great amount of attention from the differential privacy community. Often, it is the case that a successful point estimate can be derived without considering the marginal likelihood. For example, [71] shows that under mild assumptions, efficient differentially private estimators can be produced by using the *subsample-aggregate* technique.

In the case of count data, [40] show that for a wide array of utility functions, the optimal privacy mechanism can be expressed as a post-processing of a single differentially private statistic, using the *discrete Laplace* or the *geometric mechanism.* In a similar way, [7] consider all possible private hypothesis tests (that is, regardless of their test statistic) for a Bernoulli data hypothesis, and show that the uniformly most powerful differentially private test for Bernoulli data can be expressed as releasing a noisy summary statistic, and basing the test on the resulting marginal likelihood. This result indicates that while the analyst has more freedom under the query-based data access mode, especially in terms of the choice of summary statistics, the actual inference is often done in the same manner as under the dissemination-based data access model. These results are extended to other optimal inference problems for Bernoulli data, such as confidence intervals and confidence distributions, which are all products of post-processing of the same noisy summary as in [8]. These results are further extended to a broader class of differential privacy definitions, called f-DP [26], in [9].

For other problems, having control over the sensitivity of the differentially private test statistic is necessary to optimize the performance of a privatization procedure. [83] measure the sensitivity of the χ^2 statistic for contingency tables in the application to a genome-wide association study (GWAS) study. They show that the differential privacy requirement is satisfied by adding noise directly to the χ^2 statistic. Similarly, [22] and [10] measure the sensitivity of several non-parametric test statistics for comparisons of groups or paired data. Them then add noise to satisfy the differential privacy requirement, and use Monte Carlo samples or asymptotic approximations to evaluate the Type I errors. [21] show that for simple hypothesis tests, a differentially private test based on a noisy clamped likelihood ratio test achieves optimal sample complexity. [15] develop a differentially private hypothesis test for the significance of regression coefficients, based on a truncated t-statistic. [70] develop hypothesis tests and confidence intervals for linear regression, based on the Gaussian Johnson-Lindenstrauss Transform and the Analyze Gauss algorithm [35]. [3] show that for simple linear regression, noised and clamped sufficient statistics can be processed to test the significance of regression coefficients. In the locally private setting, [65] develop both frequentist and Bayesian inference on causal treatment effects in controlled experiments. So long as the private test statistic is asymptotically efficient for a null model, the methods of [14] and [6] can be used to perform approximate hypothesis tests, by using approximate co-sufficient samples. [6] demonstrate the efficacy of this differentially private testing procedure on the test of two proportions. [9] also tackle the problem of testing two proportions in the framework of f-DP, directly computing the sampling distribution by convolution via characteristic functions.

One of the fundamental statistical vocabularies for estimating and communicating uncertainty is the confidence interval. While there have been a few notable examples of producing valid confidence intervals in the dissemination model, it is more common to develop a customized query for this task. [28] develop several confidence interval mechanisms for normally distributed data, based on the parametric bootstrap, and [54] develop a finite-sample accurate confidence interval for normally distributed data, where the analysis is intricately linked with the queries asked. [27] develop the non-parametric differentially private confidence intervals for the median of one-dimensional data, using either the exponential mechanism or a private cumulative distribution function (CDF) estimator. The confidence intervals are derived using inequalities specific to the mechanisms at hand. [23] combine the techniques of subsample-aggregate, bag of little bootstraps, and *CoinPress* [18] to privately estimate a sampling distribution, producing unbiased estimates and confidence intervals with coverage that holds with high probability. Their method is based on the asymptotic theory behind the bootstrap. [78] showed that by incorportating the randomness of the nonparametric bootstrap into the privacy mechanism, multiple private releases can be given with a comparable privacy cost, while also enabling confidence intervals. [75] develop differentially private confidence intervals for output and objective perturbation mechanisms for empirical risk minimization, based on a Taylor expansion of the risk function. Besides the initial differentially private estimate, additional privacy budget must be allocated to estimate the Hessian and the covariance matrix.

There have also been a few notable works which approximate the posterior distribution based on a differentially private release, for specific models and private statistics. [16] develop a method which combines the central limit theorem and Gibbs sampling to perform private posterior inference on exponential family models, assuming that the private release is a noisy sufficient statistic. [17] extend this approach to posterior inference for Bayesian linear regression. In the setting of generalized linear models, there often does not exist a low dimensional sufficient statistic. [61] propose the use of privatized moments to approximate the differentially private posterior distribution using Normal approximations. [50] propose a general MCMC scheme that targets the private posterior distribution without approximation, applying it to a log-linear model and linear regression.

1.4 Challenges

We describe a few current challenges in the literature of statistical inference under differential privacy.

1.4.1 Disclosure risk and individual feature prediction

As discussed in Section 1.1.2, pure differential privacy bounds the probability ratios of *every* event induced by the privacy mechanism, when it is applied to a pair of neighboring databases that differ by the information of a single respondent. As a result of this construction, a differential private query bounds the probabilistic disclosure of the individual's information through the privatized query, *relative* to all the auxiliary information that the adversary may already possess of the individual as well as the other entries in the database. Using Bayesian terminology, this means that if we denote $\pi_X(x)$ as an agent's prior distribution about the individual X, and z an ϵ-differentially private query from a database containing X, we would have the guarantee that their posterior, $\pi_X(x \mid z)$, cannot change by more than $\exp(\pm\epsilon)$-fold, had z been calculated from the same database but without X.

This interpretation is called the *posterior-to-posterior semantics* of differential privacy [25, 55]. The posterior-to-posterior semantics is *counterfactual* in nature, because it invokes a comparison between two posterior probabilities based on z, one with and one without the participation of individual X, where both cannot be simultaneously true. An alternative interpretation, called the *prior-to-posterior semantics*, is concerned with the maximal change between the prior probability $\pi_X(x)$ and the posterior probability $\pi_X(x \mid z)$, also known as the (absolute) *disclosure risk* of individual X [30, 31, 62]. [12] show that under ϵ-differential privacy, this change cannot be more than $\exp(\pm\epsilon d)$-fold, where d is the maximal distance (measured according to some metric) between two permissible and connected databases from which z is obtained, so long as the prior is proper. For a non-dogmatic prior, this result sketches non-vacuous bounds on the posterior probabilities, even though in general these bounds are wider for larger databases than for smaller ones. Importantly, this means that the disclosure risk may still be high even when z is provably differentially private.

If the query release is duly privatized, when would the disclosure risk still be high? The adversary may have access to reliable auxiliary information about the individual, such as public records, common knowledge, and historical data. They may also have a good grasp of the inter-dependence between the individual and the rest of the respondents constituting the database. Taken together, this auxiliary information can be used to construct an *informative* model about X from z, in the sense that the posterior distribution $\pi_X(x \mid z)$ is concentrated on one or a small number of possible x values. Consequently, the disclosure risk of the individual can be high for certain outcomes, even if the query z is extremely, or even perfectly, private.

The relative nature of the differential privacy guarantee suggests that we cannot preclude an accurate prediction of sensitive individual features using statistical modeling based on provably private queries; see [56] for a recent example. A high predictive accuracy may be achieved if the agent knows how to construct an effective model through other means. This may happen, for example, if the target population is highly homogeneous such that a baseline prediction is already accurate for the individual. It may also happen if, even without the private query input, the prior distribution $\pi_X(x)$ is already highly informative.

While an informative posterior or predictive probability of the individual's features does not constitute a violation of differential privacy, they are, in a sense, still a threat to the individual's confidentiality due to the elevated disclosure risk. Unfortunately, privacy mechanisms cannot help limit the impact of information sources that already exist, or be derived from those that are. Nevertheless, the assessment of the actual disclosure risk, rather than the privacy guarantee itself, can be of vital importance for many use cases. We refer the reader to Chapter [disclosure-risk] of this book for an extended discussion.

1.4.2 Invariants and logical constraints

Official statistical agencies curate data in accordance with legal and policy mandates. Sometimes, this means that certain aspects of the data products must be released exactly without privacy error, or that they must satisfy certain logical constraints to maintain internal consistency. For example, to release the 2020 U.S. Decennial Census under differential privacy, the Census Bureau observes a set of *invariants* [4], such as population totals at the state level, total housing units and group quarter facilities at the block level, and so on. The population count of a higher geographic level must be equal to those tabulated from its lower-level constituents. In addition, all privatized counts must be non-negative. If the counts were simply noise-infused and tabulated, the resulting data product may well be inconsistent with the required invariants and constraints. To enforce these invariants, something must be done after, or in lieu of, the straightforward noise infusion mechanism.

Unfortunately, differential privacy and invariants do not mingle well. A commonly used method to impose invariants on a differentially private noisy query is post-processing using distance minimization, as employed by the Census DAS TopDown algorithm [1]. The underlying optimization procedure is highly data-dependent, and the statistical implication is two-fold. First, this approach may introduce systematic bias into the query output that is difficult to quantify. For example, it has been observed that the TopDown algorithm tends to associate larger counts with negative errors, whereas smaller counts with positive errors, when the total count is held as invariant [84]. Second, the analytically intractable and computationally complex procedure sacrifices the transparent probabilistic description of the privacy mechanism, leading to challenges in the statistical inference task based on these data, as Section 1.2.1 discusses.

The literature has seen some efforts to design privacy mechanisms that integrate invariants and constraints as part of the process. [13, 48] discuss the design of privacy mechanisms that maintain internal logical consistency in the multi-dimensional output. [39] and [24] discuss the design of privacy mechanisms that obey linear invariants that are exact statistics of the confidential data, warning about the limited extent to which they protect privacy relative to the classic notion of differential privacy. [69] instead use the Pufferfish framework [57] as an alternative method of quantifying partially private data. From a statistical point of view, these designs better preserve the probabilistic transparency of the generative process of the query release, facilitating downstream statistical analysis and the reliable quantification of uncertainty.

1.4.3 Unbounded query sensitivity and clamping

One of the simplest statistics to privatize is a count, which is a sum of binary data points. Specifically, each individual's contribution to the sum is either one or zero, hence the *sensitivity* of the counting query is naturally bounded. In this case, directly adding noise with an appropriate scale to the query achieves formal privacy. This property extends to other data with a bounded domain, such as in Example 1. In these cases, the privatized output is *unbiased*, since the additive noise has an expectation of zero, and the original statistic is not altered in any way.

However, many statistics and data structures do not have a bounded sensitivity. Consider for example the sample average, standard deviation, or regression coefficients calculated based on real-valued data. Theoretically, it is possible for an individual's information to contribute an infinitely large change to the value of the statistic, hence no additive noise with a finite scale would suffice for differential privacy. In any practical setting, however, we do not really expect to observe data values that are infinitely large. In survey statistics, *top-coding* or the censoring of data values above a certain upper bound, is frequently practiced. For example, the Current Population Survey top-codes an individual's hourly earnings such that the annualized wage lies between \$0 and \$150k inclusive [72]. All observations that exceed the upper limit are adjusted downward to \$150k. In differential privacy, the same procedure is traditionally employed as part of data pre-processing, known as *clamping*, to render all data values inside a bounded interval chosen in advance. For a dataset of size n that is top-coded (or clamped) as above, the average annualized wage would have a sensitivity of \$150k/n.

Because clamping systematically alters the statistic calculated from the underlying confidential data, it may introduce bias into the differentially private estimate, even if the additive noise itself is unbiased. In fact, there exists a bias-variance trade-off in the choice of the clamping thresholds. Since noise is scaled proportionally to the length of the clamping interval, choosing clamping thresholds that are too wide introduces excessive noise. On the other hand, as the thresholds become narrower, more data are systematically altered, and the bias becomes larger.

Besides adding noise directly, there are more sophisticated privacy methods that work by privately minimizing a loss function, such as the log-likelihood function. These methods include the exponential mechanism or objective perturbation. In this setting, the sensitivity of the loss function may still be unbounded, unless the data is first clamped. A workaround is to modify the score by applying a bounded function, as is often used in robust statistics. While this can mitigate the dramatic effects of clamping, altering the score functions often still results in some bias.

A clever approach to the problem of choosing the clamping threshold is *CoinPress*, proposed in [18]. This algorithm follows a two-step procedure. First, the method uses a proportion of the privacy budget to iteratively estimate the range of the data before clamping. It then uses the estimated range to choose a well-informed clamp before adding noise. Similar ideas have been used in [54] and [71], where they use a quantile approach to approximately estimate the support of the data before clamping and adding noise. This two-step approach can mitigate the worst effects of the clamping, but does not eliminate the bias completely.

Unfortunately, except in the simplest of settings, some form of bias is typically introduced by the differential privacy procedure, either through clamping the data or altering the loss function. There may be room for improvement on the mechanism design front. On the other hand, the bias can also be addressed using statistical post-processing techniques, such as those discussed in Section 1.2.3. For example, parametric bootstrap and other simulation-based inference methods can reduce the bias of an estimator to $O(0)$ (that is $O(n^p)$ for all p) [45, 46]. [11] show that simulation-based inference can account for the bias due to clamping, enabling valid confidence intervals and hypothesis tests. The marginal likelihood can also be used to derive the maximum likelihood estimator based on the private output. For example, [53] use this approach to derive the differentially private MLE for the β-model on network data.

1.4.4 Identifiability, Estimability, and Consistency

In practical data analysis, most statistical models that we employ are *identifiable* by design. A model is called identifiable, if $f(\cdot \mid \theta) = f(\cdot \mid \theta')$ implies that $\theta = \theta'$. In other words, different parameters result in different models. If a model is identifiable, then given enough data, it is possible to perfectly identify the true parameters that generated the model. On the other hand, if a model is non-identifiable, then no matter how much data we have, we will never learn the true parameters.

There is an interesting phenomenon in privacy that even supposing the original model $f(x \mid \theta)$ is identifiable, and the privacy mechanism $\eta(z \mid x)$ is known, it may be that the marginal model for Z, $p(z \mid \theta)$ is *not* identifiable. Under the dissemination mode, this can easily happen if none of the privatized statistics are informative about the particular parameters of interest. For example, if the model aims to capture the correlation between two groups, the parameter cannot be well-identified if only the marginal count for either group is released. Under the query-based mode of access, it still may be a challenge to choose an appropriate privatized statistics to ensure that the model is identifiable.

A related, but distinct, concept to identifiability is *estimability.* A parameter is estimable if there exists an unbiased estimator for it. There exists non-identifiable models which have estimable parameters. For example in classic linear regression, if two or more of the predictor variables are collinear, the associated regression coefficients become aliased and non-identifiable. However, functions of the coefficients, such as the fitted values of the response variables, may still be estimable. On the other hand, there also exists identifiable models which have inestimable parameters. For example, suppose that the output of a privacy mechanism is an estimate of the regression coefficients. The distribution of this differentially private estimator depends on σ^2, the scale of the idiosyncratic error in the regression model. However, based only on a single privatized regression coefficient estimator, there is no unbiased estimator for σ^2. In this case, σ^2 is identifiable but not estimable.

We have already discussed the possibility of bias and inconsistency in differentially private estimators, which may arise with invariants, constraints, and clamping. The inconsistency of a particular differentially private estimator should not be confused with the lack of estimability of the parameter. If a parameter is not estimable, there may not exist a consistent differentially private estimator for it. However, it may still be the case that some function of the differentially private estimator gives a consistent estimate of the parameter. For example, when using truncation, the truncation parameter must be carefully incorporated in the asymptotic regime to ensure consistency [59]. So long as the distribution of the inconsistent estimator changes as the parameter varies, a consistent estimator may be possible [45].

1.5 Discussion

Statistics is the mathematical science of understanding uncertainty, and incorporating uncertainty into the expression of scientific knowledge, whatever the source may be. If there is uncertainty in the data collection method, missing data, or if the observed data are contaminated by measurement errors, statisticians have found it crucial to include these aspects of uncertainty into the model, and propagate them through the statistical analyses. The randomness introduced by a privacy mechanism is no different than these other sources of error. Except that in the case of privacy, we are fortunate enough to be able to know the exact conditional distribution of the privatized output given the original dataset, a property that is usually not enjoyed by measurement error and missing data problems. It is the transparency of the probabilistic characteristics of the privacy problem that allows us to express the marginal likelihood of the privatized query precisely as a convolution, integrating over the latent space of databases. While this likelihood is often intractable, there are already a suite of computational and theoretical tools to perform valid statistical inference based on the privatized output.

There is plenty of room for future work. Many current inference techniques are tailored to a specific model, privacy mechanism, or statistical task. The development of general purpose methods – computationally and practically accessible ones – can greatly help data users to perform valid statistical inference. They are important to ensure that differentially private data releases are properly understood.

Bibliography

[1] John M Abowd, Robert Ashmead, Garfinkel Simson, Daniel Kifer, Philip Leclerc, Ashwin Machanavajjhala, and William Sexton. Census topdown: Differentially private data, incremental schemas, and consistency with public knowledge. Technical report, US Census Bureau, 2019.

[2] John M Abowd and Ian M Schmutte. An economic analysis of privacy protection and statistical accuracy as social choices. *American Economic Review*, 109(1):171–202, 2019.

[3] Daniel Alabi and Salil Vadhan. Hypothesis testing for differentially private linear regression. *Advances in Neural Information Processing Systems*, 35:14196–14209, 2022.

[4] Robert Ashmead, Daniel Kifer, Philip Leclerc, Ashwin Machanavajjhala, and William Sexton. Effective privacy after adjusting for invariants with applications to the 2020 census. Technical report, US Census Bureau, 2019.

[5] Marco Avella-Medina. Privacy-preserving parametric inference: a case for robust statistics. *Journal of the American Statistical Association*, 116(534):969–983, 2021.

[6] Jordan Awan and Zhanrui Cai. One step to efficient synthetic data. *arXiv e-prints*, pages arXiv–2006, 2020.

[7] Jordan Awan and Aleksandra Slavković. Differentially private uniformly most powerful tests for binomial data. *Advances in Neural Information Processing Systems*, 31:4208–4218, 2018.

[8] Jordan Awan and Aleksandra Slavković. Differentially private inference for binomial data. *Journal of Privacy and Confidentiality*, 10(1):1–40, 2020.

[9] Jordan Awan and Salil Vadhan. Canonical noise distributions and private hypothesis tests. *arXiv preprint arXiv:2108.04303*, 2021.

[10] Jordan Awan and Yue Wang. Differentially private Kolmogorov-Smirnov-type tests. *arXiv preprint arXiv:2208.06236*, 2022.

[11] Jordan Awan and Zhanyu Wang. Simulation-based, finite-sample inference for privatized data. *arXiv preprint arXiv:2303.05328*, 2023.

[12] James Bailie and Ruobin Gong. Differential privacy: General inferential limits via intervals of measures. In *Proceedings of the Thirteenth International Symposium on Imprecise Probability: Theories and Applications (ISIPTA'23)*, 2023.

[13] Boaz Barak, Kamalika Chaudhuri, Cynthia Dwork, Satyen Kale, Frank McSherry, and Kunal Talwar. Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In *Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 273–282, 2007.

[14] Rina Foygel Barber and Lucas Janson. Testing goodness-of-fit and conditional independence with approximate co-sufficient sampling. *The Annals of Statistics*, 50(5):2514–2544, 2022.

[15] Andrés F Barrientos, Jerome P Reiter, Ashwin Machanavajjhala, and Yan Chen. Differentially private significance tests for regression coefficients. *Journal of Computational and Graphical Statistics*, 28(2):440–453, 2019.

[16] Garrett Bernstein and Daniel R Sheldon. Differentially private Bayesian inference for exponential families. *Advances in Neural Information Processing Systems*, 31:2919–2929, 2018.

[17] Garrett Bernstein and Daniel R Sheldon. Differentially private Bayesian linear regression. *Advances in Neural Information Processing Systems*, 32:525–535, 2019.

[18] Sourav Biswas, Yihe Dong, Gautam Kamath, and Jonathan Ullman. Coinpress: Practical private mean and covariance estimation. *Advances in Neural Information Processing Systems*, 33, 2020.

[19] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pages 635–658. Springer, 2016.

[20] T Tony Cai, Yichen Wang, and Linjun Zhang. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *The Annals of Statistics*, 49(5):2825–2850, 2021.

[21] Clément L Canonne, Gautam Kamath, Audra McMillan, Adam Smith, and Jonathan Ullman. The structure of optimal private tests for simple hypotheses. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 310–321, 2019.

[22] Simon Couch, Zeki Kazan, Kaiyan Shi, Andrew Bray, and Adam Groce. Differentially private nonparametric hypothesis testing. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 737–751, 2019.

[23] Christian Covington, Xi He, James Honaker, and Gautam Kamath. Unbiased statistical estimation and valid confidence intervals under differential privacy. *arXiv preprint arXiv:2110.14465*, 2021.

[24] Prathamesh Dharangutte, Jie Gao, Ruobin Gong, and Fang-Yi Yu. Integer subspace differential privacy. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-23)*, 2023.

[25] Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 202–210, 2003.

[26] Jinshuo Dong, Aaron Roth, and Weijie J Su. Gaussian differential privacy. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):3–37, 2022.

[27] Jörg Drechsler, Ira Globus-Harris, Audra Mcmillan, Jayshree Sarathy, and Adam Smith. Nonparametric differentially private confidence intervals for the median. *Journal of Survey Statistics and Methodology*, 10(3):804–829, 2022.

[28] Wenxin Du, Canyon Foot, Monica Moniot, Andrew Bray, and Adam Groce. Differentially private confidence intervals. *arXiv preprint arXiv:2001.02285*, 2020.

[29] John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 429–438. IEEE, 2013.

[30] George T Duncan and Diane Lambert. Disclosure-limited data dissemination. *Journal of the American Statistical Association*, 81(393):10–18, 1986.

[31] George T Duncan and Diane Lambert. The risk of disclosure for microdata. *Journal of Business & Economic Statistics*, 7(2):207–217, 1989.

[32] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. Preserving statistical validity in adaptive data analysis. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 117–126, 2015.

[33] Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 371–380, 2009.

[34] Cynthia Dwork and Guy N Rothblum. Concentrated differential privacy. *arXiv preprint arXiv:1603.01887*, 2016.

[35] Cynthia Dwork, Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Analyze gauss: optimal bounds for privacy-preserving principal component analysis. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 11–20, 2014.

[36] Cecilia Ferrando, Shufan Wang, and Daniel Sheldon. Parametric bootstrap for differentially private confidence intervals. In *International Conference on Artificial Intelligence and Statistics*, pages 1598–1618. PMLR, 2022.

[37] Marco Gaboardi, Hyun Lim, Ryan Rogers, and Salil Vadhan. Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing. In *International Conference on Machine Learning*, pages 2111–2120. PMLR, 2016.

[38] Marco Gaboardi and Ryan Rogers. Local private hypothesis testing: Chi-square tests. In *International Conference on Machine Learning*, pages 1626–1635. PMLR, 2018.

[39] Jie Gao, Ruobin Gong, and Fang-Yi Yu. Subspace differential privacy. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-22)*, 36(4):3986–3995, 2022. doi:10.1609/aaai.v36i4.20315.

[40] Arpita Ghosh, Tim Roughgarden, and Mukund Sundararajan. Universally utility-maximizing privacy mechanisms. *SIAM Journal on Computing*, 41(6):1673–1693, 2012.

[41] Harvey Goldstein and Natalie Shlomo. A probabilistic procedure for anonymisation, for assessing the risk of re-identification and for the analysis of perturbed data sets. *Journal of Official Statistics*, 36(1):89–115, 2020.

[42] Ruobin Gong. Exact inference with approximate computation for differentially private data via perturbations. *Journal of Privacy and Confidentiality*, 12(2), 2022.

[43] Ruobin Gong. Transparent privacy is principled privacy. *Harvard Data Science Review (Special Issue 2)*, 2022. doi:10.1162/99608f92.b5d3faaa.

[44] J M Gouweleeuw, Peter Kooiman, and PP De Wolf. Post randomisation for statistical disclosure control: Theory and implementation. *Journal of official Statistics*, 14(4):463, 1998.

[45] Stéphane Guerrier, Elise Dupuis-Lozeron, Yanyuan Ma, and Maria-Pia Victoria-Feser. Simulation-based bias correction methods for complex models. *Journal of the American Statistical Association*, 114(525):146–157, 2019.

[46] Stéphane Guerrier, Mucyo Karemera, Samuel Orso, and Maria-Pia Victoria-Feser. Asymptotically optimal bias reduction for parametric models. *arXiv preprint arXiv:2002.08757*, 2020.

[47] Michael B Hawes. Implementing differential privacy: Seven lessons from the 2020 United States Census. *Harvard Data Science Review*, 2(2), 2020.

[48] Michael Hay, Vibhor Rastogi, Gerome Miklau, and Dan Suciu. Boosting the accuracy of differentially private histograms through consistency. *Proceedings of the VLDB Endowment*, 3(1), 2010.

[49] V Joseph Hotz, Christopher R Bollinger, Tatiana Komarova, Charles F Manski, Robert A Moffitt, Denis Nekipelov, Aaron Sojourner, and Bruce D Spencer. Balancing data privacy and usability in the federal statistical system. *Proceedings of the National Academy of Sciences*, 119(31):e2104906119, 2022.

[50] Nianqiao Ju, Jordan Awan, Ruobin Gong, and Vinayak Rao. Data augmentation MCMC for Bayesian inference from privatized data. *Advances in Neural Information Processing Systems*, 35:12732–12743, 2022.

[51] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. In *International conference on machine learning*, pages 1376–1385. PMLR, 2015.

[52] Vishesh Karwa, Dan Kifer, and Aleksandra B Slavković. Private posterior distributions from variational approximations. *arXiv preprint arXiv:1511.07896*, 2015.

[53] Vishesh Karwa and Aleksandra Slavković. Inference using noisy degrees: Differentially private β-model and synthetic graphs. *The Annals of Statistics*, 44(1):87–112, 2016.

[54] Vishesh Karwa and Salil Vadhan. Finite sample differentially private confidence intervals. In *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

[55] Shiva P Kasiviswanathan and Adam Smith. On the'semantics' of differential privacy: A Bayesian formulation. *Journal of Privacy and Confidentiality*, 6(1), 2014.

[56] Christopher T Kenny, Shiro Kuriwaki, Cory McCartan, Evan TR Rosenman, Tyler Simko, and Kosuke Imai. The use of differential privacy for census data and its impact on redistricting: The case of the 2020 us census. *Science Advances*, 7(41):eabk3283, 2021.

[57] Daniel Kifer and Ashwin Machanavajjhala. Pufferfish: A framework for mathematical privacy definitions. *ACM Transactions on Database Systems (TODS)*, 39(1):1–36, 2014.

[58] Daniel Kifer, Adam Smith, and Abhradeep Thakurta. Private convex empirical risk minimization and high-dimensional regression. In *Conference on Learning Theory*, pages 25–1. JMLR Workshop and Conference Proceedings, 2012.

[59] Tatiana Komarova and Denis Nekipelov. Identification and formal privacy guarantees. *Available at SSRN*, 2020.

[60] Tejas Kulkarni, Joonas Jälkö, Samuel Kaski, and Antti Honkela. Locally differentially private Bayesian inference. *arXiv preprint arXiv:2110.14426*, 2021.

[61] Tejas Kulkarni, Joonas Jälkö, Antti Koskela, Samuel Kaski, and Antti Honkela. Differentially private Bayesian inference for generalized linear models. In *International Conference on Machine Learning*, pages 5838–5849. PMLR, 2021.

[62] David McClure and Jerome P Reiter. Differential privacy and statistical disclosure risk measures: An investigation with binary synthetic data. *Trans. Data Priv.*, 5(3):535–552, 2012.

[63] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103. IEEE, 2007.

[64] Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pages 263–275. IEEE, 2017.

[65] Yuki Ohnishi and Jordan Awan. Locally private causal inference. *arXiv preprint arXiv:2301.01616*, 2023.

[66] Silvia Polettini and Serena Arima. Small area estimation with covariates perturbed for disclosure limitation. *Statistica*, 75(1):57–72, 2015.

[67] Nancy Potok. Deep policy learning: Opportunities and challenges from the evidence act. *Harvard Data Science Review*, 1(2), 2019.

[68] Matthew Reimherr and Jordan Awan. Elliptical perturbations for differential privacy. *Advances in Neural Information Processing Systems*, 32, 2019.

[69] Jeremy Seeman, Aleksandra Slavkovic, and Matthew Reimherr. A formal privacy framework for partially private data. *arXiv preprint arXiv:2204.01102*, 2022.

[70] Or Sheffet. Differentially private ordinary least squares. In *International Conference on Machine Learning*, pages 3105–3114. PMLR, 2017.

[71] Adam Smith. Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 813–822, 2011.

[72] U. S. Census Bureau. Current population survey, October 2021, school enrollment technical supplement documentation, attachment 12, October 2021.

[73] U.S. Census Bureau. Now online: New disclosure avoidance system update meets or exceeds redistricting accuracy targets, April 2021. Retrieved from https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance/2020-das-updates/2021-04-28.html on Jan 24, 2022.

[74] Duy Vu and Aleksandra Slavkovic. Differential privacy for clinical trial data: Preliminary evaluations. In *2009 IEEE International Conference on Data Mining Workshops*, pages 138–143. IEEE, 2009.

[75] Yue Wang, Daniel Kifer, and Jaewoo Lee. Differentially private confidence intervals for empirical risk minimization. *Journal of Privacy and Confidentiality*, 9(1), 2019.

[76] Yue Wang, Daniel Kifer, Jaewoo Lee, and Vishesh Karwa. Statistical approximating distributions under differential privacy. *Journal of Privacy and Confidentiality*, 8(1), 2018.

[77] Yue Wang, Jaewoo Lee, and Daniel Kifer. Revisiting differentially private hypothesis tests for categorical data. *arXiv preprint arXiv:1511.03376*, 2015.

[78] Zhanyu Wang, Guang Cheng, and Jordan Awan. Differentially private bootstrap: New privacy analysis and inference strategies. *arXiv preprint arXiv:2210.06140*, 2022.

[79] Stanley L Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

[80] Larry Wasserman and Shuheng Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010.

[81] Oliver Williams and Frank McSherry. Probabilistic inference and differential privacy. *Advances in Neural Information Processing Systems*, 23:2451–2459, 2010.

[82] Yong Ming Jeffrey Woo and Aleksandra Slavkovic. Generalised linear models with variables subject to post randomization method. *Statistica Applicata-Italian Journal of Applied Statistics*, 24(1):29–56, 2015.

[83] Fei Yu, Stephen E Fienberg, Aleksandra B Slavković, and Caroline Uhler. Scalable privacy-preserving data sharing methodology for genome-wide association studies. *Journal of biomedical informatics*, 50:133–141, 2014.

[84] Keyu Zhu, Pascal Van Hentenryck, and Ferdinando Fioretto. Bias and variance of post-processing in differential privacy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11177–11184, 2021.