Privacy Tools
for Sharing Research
Data

*A National Science Foundation
Secure and Trustworthy Cyberspace
Project*

*with additional support from the Sloan Foundation and
Google, Inc.*

Differential Privacy in CDFs

Daniel Muise, Kobbi Nissim
Center for Research on Computation and Society
Harvard University
April 2016

Work supported by NSF grant CNS-1237235, grants from the Sloan Foundation, and a Simons Investigator grant to Salil Vadhan.

State of the Presentation

This version of the presentation has been made available for general commentary and review, and should not be regarded as an absolute final product.

Acknowledgments

We thank Mark Bun, Vishesh Karwa, and Salil Vadhan for insightful commentary, and Georgios Kellaris whose implementations helped create several images in this presentation.

Differential Privacy in CDFs

The ultimate aim of this presentation is to familiarize social scientists with the errors introduced by differential privacy (DP), and to explain how to manage DP's random noise.

In this document, we explain the effect of random noise introduced in DP-computations by making analogies to sampling error. We focus on the case of cumulative density functions (CDFs) and histograms.

Other supporting documents can be found at the project's webpage: http://privacytools.seas.harvard.edu

Table of Contents

Section 1

WHAT IS DIFFERENTIAL PRIVACY? A BRIEF INTRODUCTION.

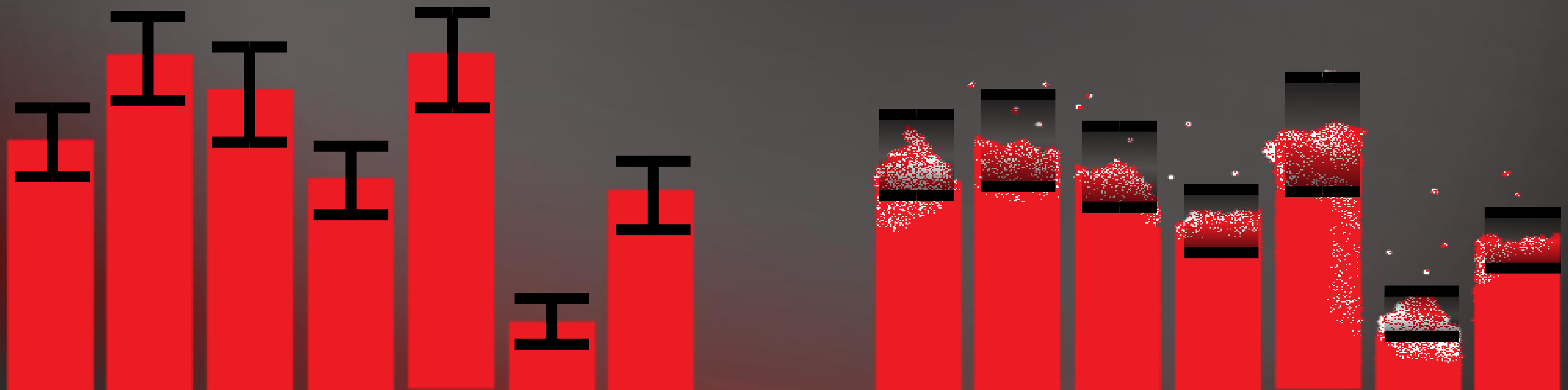
Differential Privacy

Differential Privacy allows us to look at data through a clouded lens: it allows us to see trends in data while hiding the specifics of individual records.

Our aim: Research data presently hidden from the public (due to sensitivity and participants' privacy concerns) can be made safely available in an 'obscured' form.

Differential Privacy

This 'obscuring' of data is done with the addition of random noise in a <u>controllable</u>, mathematical way so as to satisfy a precise privacy goal.



Differential Privacy

The following slides convey how using differentially private statistics can be intuitive, as the introduction of random noise is in many ways similar to the inherent randomness of sampling error.

Just as inferences from sample data provide useful insight into population properties, inferences from differentially private data can suffice for many purposes.

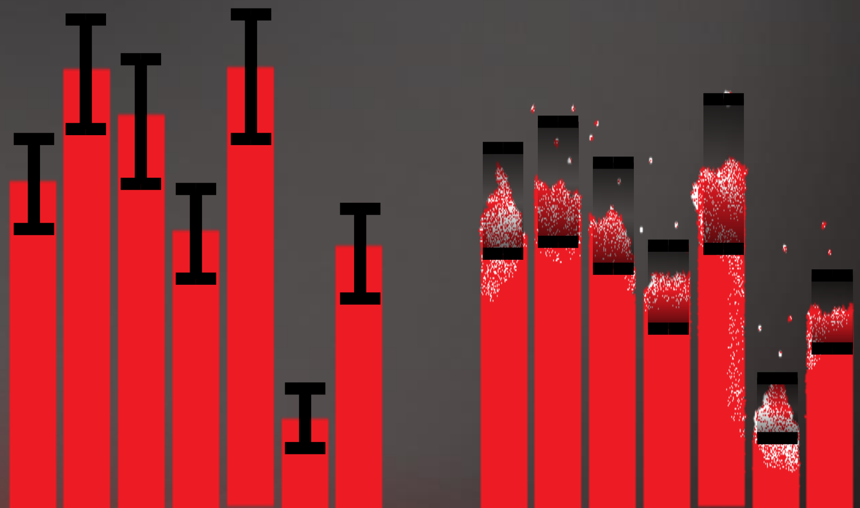
Sampling Error

- In general, sampling error occurs when small datasets are used to make inferences about the populations they are drawn from.
- The accuracy of a sample-based estimate depends on the sample size and representativeness. Absolute precision would require sampling an entire population, but past a certain sample size the diminishing gains in precision become negligible.

Differential Privacy & Sampling Error

Similarities:

- Both may adversely influence conclusions drawn from data.
- In both, errors are more severe with smaller sample size ***n*.**
- Magnitude of errors can be estimated.



Section 2.1

SAMPLING ERROR AND INFERENCE IN SOCIAL SCIENCE

Means

Histograms

CDFs

Statistics as Approximation

To broach the subject of differential privacy, and why addition of random noise is not so strange, we must first become aware of two things:

Statistics is a method of estimation,

and

Statistical estimation copes with the inherent randomness of sampling error.

Statistics as Approximation

When used correctly, statistics is effective despite these two facts. Statistics have predicted important events, trends, and catastrophes. Statisticians are paid millions by stock brokers who then make millions themselves. We base our economy and national future on statistical data.

These are basic examples of how statistical computations are often accurate and useful despite being approximations.

Case Example

The following slides will provide a simple example of sampling error as a product of randomness.

They will explain how and why it arises, and how the same methodology can produce different results, due to randomness.

Case example: Mean Number of Social Groups

Consider a political scientist, Neil. Neil is interested in civil society, so he's studying the relationship between citizens' social engagement and the outcome of local elections.

For his current work, he needs to find the average number of social groups that each person identifies with in each district.

To find that exact average, Neil would need to track down and survey everybody in the district, and ask them about their social life. Neil cannot do this, as it would cost too much time and money. Thus, Neil must use statistics to estimate that average.

Sampling error

Currently, Neil is studying District X. For simplicity's sake, let's say District X has 90 people in it. In the chart below, each box represents a person.

Number of Social Groups each Person Identifies with in District X

3	5	3	5	3	3	4	5	4	5	3	5	5	5	3	0	8	6	8	6	9	6	4	0	4	7	6	8	7	2
1	0	5	5	3	3	5	7	0	3	2	3	1	3	0	1	5	7	2	4	6	9	7	0	9	9	4	5	9	1
4	4	2	6	7	9	3	0	9	3	5	6	3	2	9	5	8	6	7	5	3	4	3	9	5	0	3	3	2	0

Here we see the average number of social groups is <u>4.4</u>, found by summing the observations and dividing by ***n*** (90).

As we said, Neil doesn't have the time and money to survey all 90 people. He can only survey 18. Randomly selecting 18, he <u>estimates</u> an average of <u>2.3 groups per person!</u>

															0								0						2
1	0								3			1							4						9				
		2					0						3				6			3				5	0			2	0

Sampling error

Number of Social Groups each Person Identifies with in District X

3	5	3	5	3	3	4	5	4	5	3	5	5	5	3	0	8	6	8	6	9	6	4	0	4	7	6	8	7	2
1	0	5	5	3	3	5	7	0	3	2	3	1	2	0	1	5	7	2	4	6	9	7	0	9	9	4	5	9	1
4	4	2	6	7	9	3	0	9	3	5	6	3	3	9	5	8	6	7	5	3	4	3	9	5	0	3	3	2	0

Neil's three colleagues don't trust his answer, and so they each do their own random sampling of District X.

Means:

3							5						5							9	6				7			7		
							7				3					5					9				9					**5.72**
	4																	7		3			9	5	0					

					3					3			5					8									8			
							7				3					5							0	9						**4.33**
							0							9	5					3	4				0	3	3			

															0							4				6			2	
1		5						0	3									2			9		0							**3.22**
				7						5				9						3					0			2	0	

Sampling error

Here we see each of the researchers' results.

Researcher	Sample Size	Mean number of groups that each person identifies with in District X		Error from Sampling
		Estimated	True	
Neil	18	2.30	4.4	-2.10
Colleague 1	18	5.72	4.4	1.32
Colleague 2	18	4.33	4.4	-0.07
Colleague 3	18	3.22	4.4	-1.18

Their different results are a testament to the randomness of selecting which 18 people in District X to survey. None of the colleagues knew better or worse which sample of 18 would be more representative of the whole district, and so in their eyes all answers are equally possible.

Sampling error

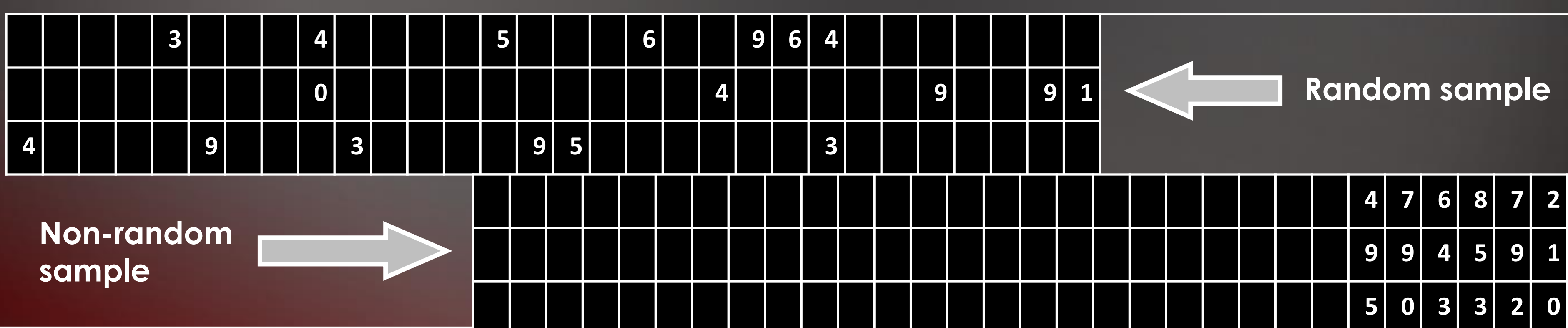
As the example demonstrates, sampling error occurs when we estimate properties of data we can't fully access.

Randomness is key here. We don't know for sure how representative our sample is. Neil couldn't choose which observations he would collect, and doesn't know how well his observations align with the rest of District X.

Estimation techniques and norms in social science have developed to accommodate sampling error.

Random Sampling

It's important to note that randomness is used as a tool in statistics, to avoid overrepresentation of certain values, and thus minimize sampling error. Neil and his colleagues were right to sample *randomly*, instead of only surveying wealthy neighborhoods, or elderly citizens, for example. However, sampling error is present regardless, even if *random* sampling helps ensure that it stays low.



Sampling error

The findings of Neil and his colleagues clearly showed the effects of sampling, but the district size of *90* is unrealistically small, and the relative sample size, 20%, is unrealistically high. Both were exaggerated for demonstration.

In the following pages, we'll work with larger datasets, ones that can only be visualized with graphs and histograms.

When looking at the following pages, keep in mind that we're only showing <u>one possible sample</u> at each sample size. Due to the randomness of sampling, for any large dataset, there can be thousands, or even millions and billions, of possible samples.

Sampling error

We will use the next example to show that better results can be gained by altering our sample size. We like to think of sample size as a knob we can turn. With a larger *n*, we can get more accurate results. To see it a different way, turning the "knob" of n lowers the influence of randomness.

INNACCURATE

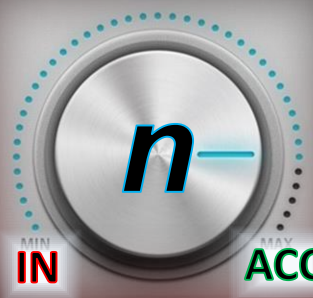
larger randomness influence

larger sample size

ACCURATE

Later on, we will see that differential privacy also has this "n knob." It also has a special knob to itself, ε (epsilon).

larger randomness influence



larger sample size

more random noise, more privacy

Less random noise, less privacy

Section 2.2

SAMPLING ERROR AND INFERENCE IN SOCIAL SCIENCE

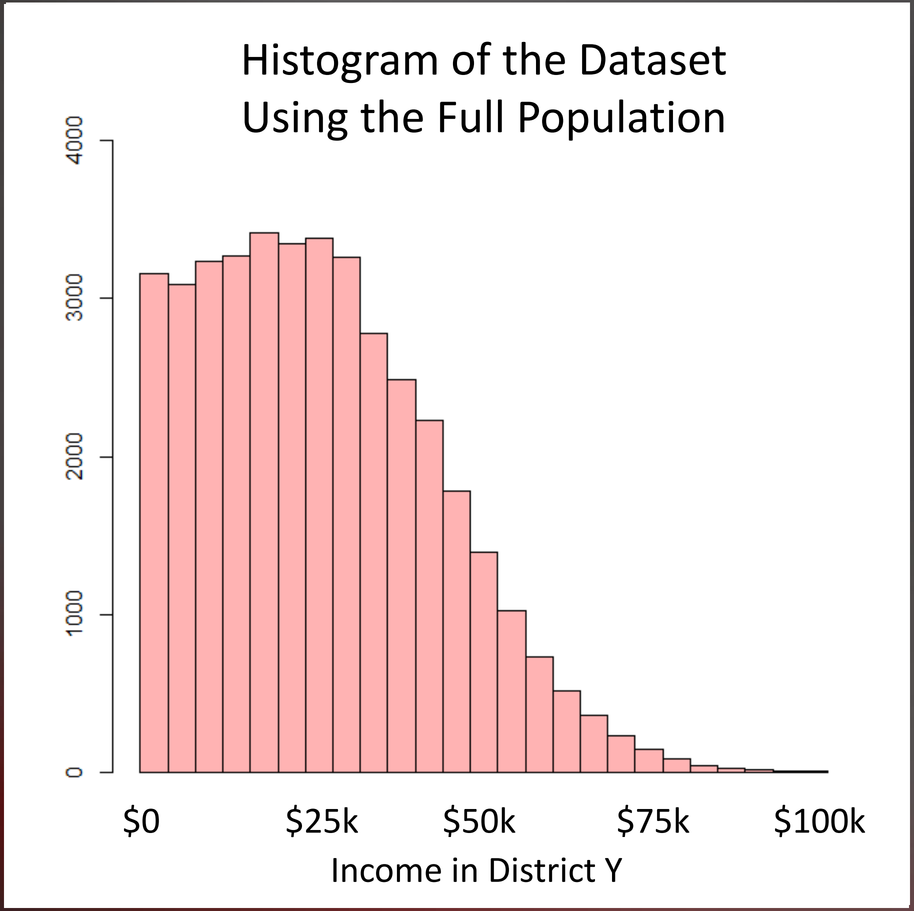
Means

Histograms

CDFs

Sampling error in practice

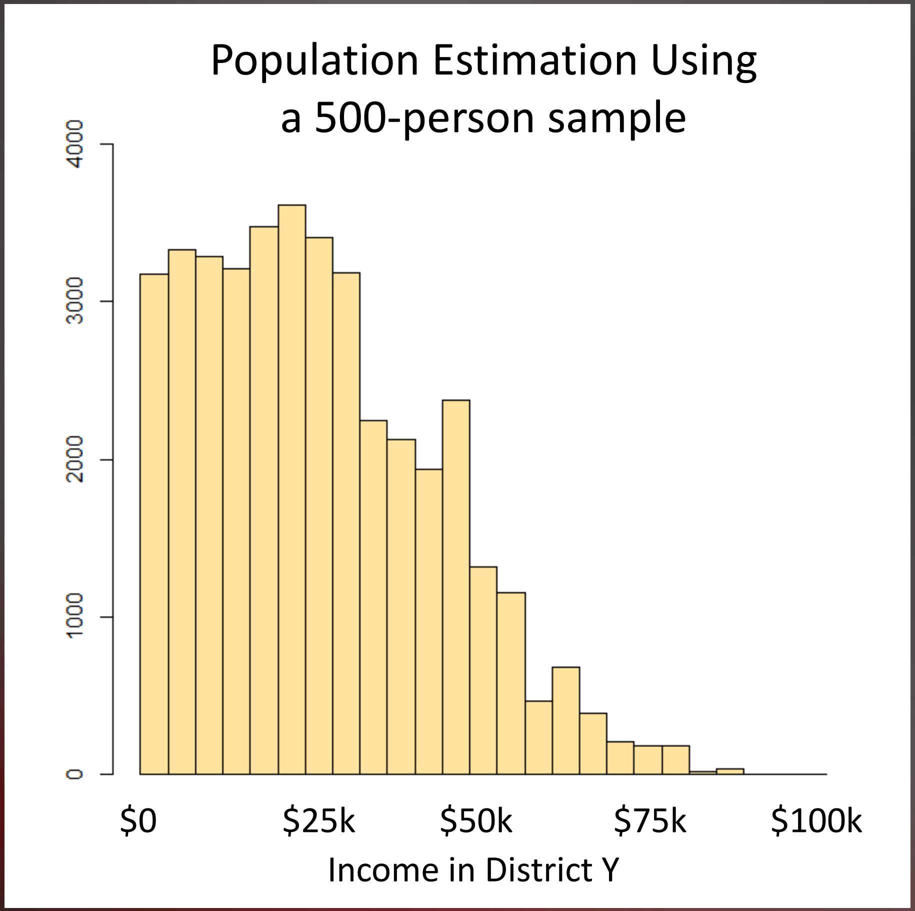
Gertrude is an economist, and she needs to know the mean income in District Y, which has 40,000 people. District Y's income distribution is plotted below, with the mean being $26,000.



As with Neil, Gertrude doesn't have the resources to survey all 40,000 people in the district, so she'll never know this true mean or distribution. Instead, she takes a random sample of citizens, and estimates as best she can.

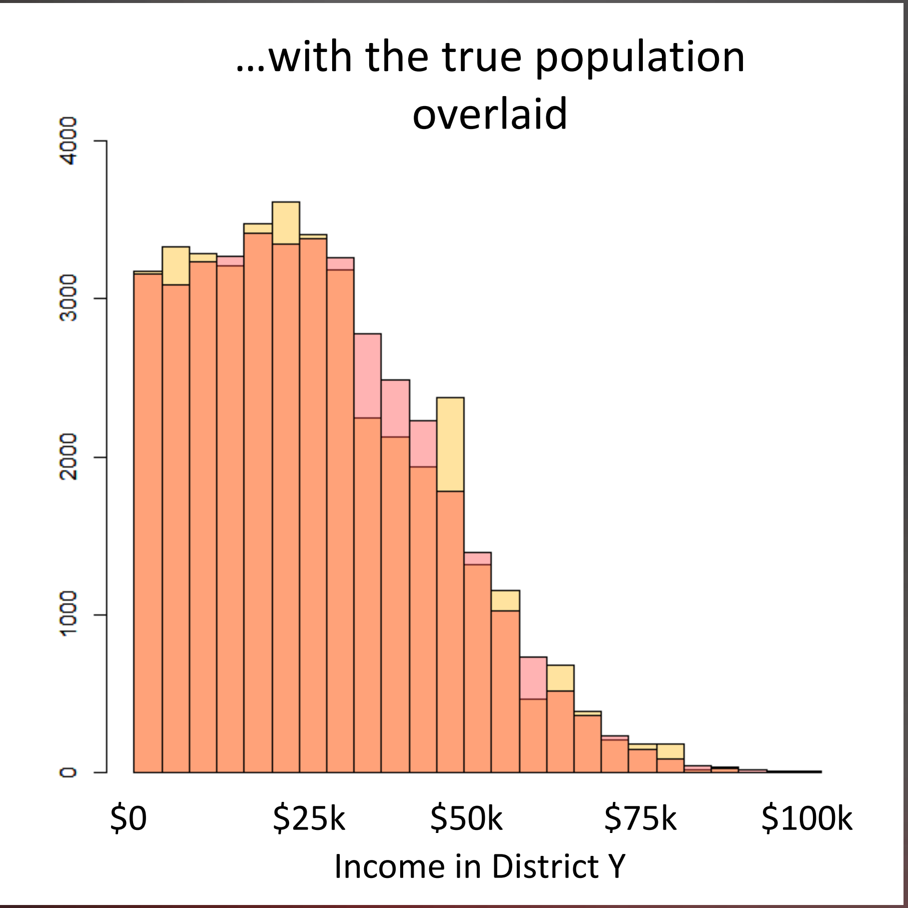
Sampling error in a histogram

Given her time and funding constraints, Gertrude manages to survey 500 people.



Since Gertrude is using this sample to estimate properties of the whole District Y population, each observation represents 80 people in the population. Using this basic method, this graph summarizes our prediction for the population, based on the sample.

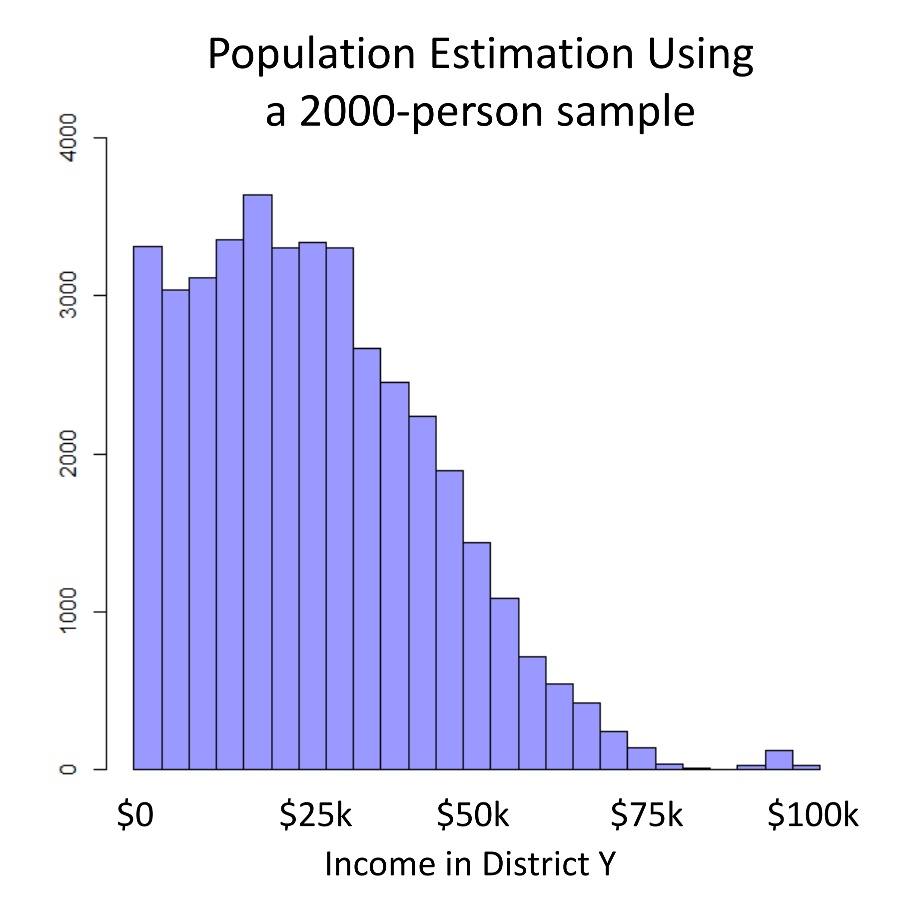
Sampling error in a histogram

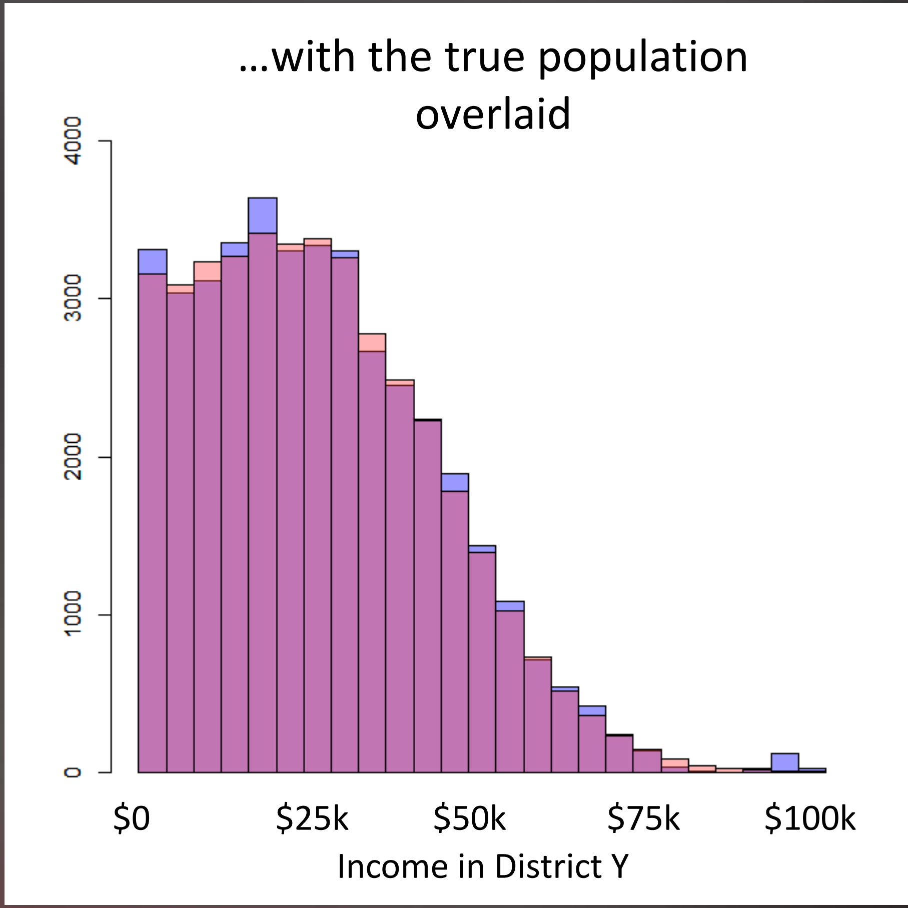


Overlaying the graph based on Gertrude's sample (**yellow**) with the graph of the real population (**red**), we can use this resulting (**orange**) graph to see small differences arising. In bins with red tops, the sampling underestimated the frequency of that bin's value, while the opposite is true of bins with yellow tops.

Sampling error in a histogram

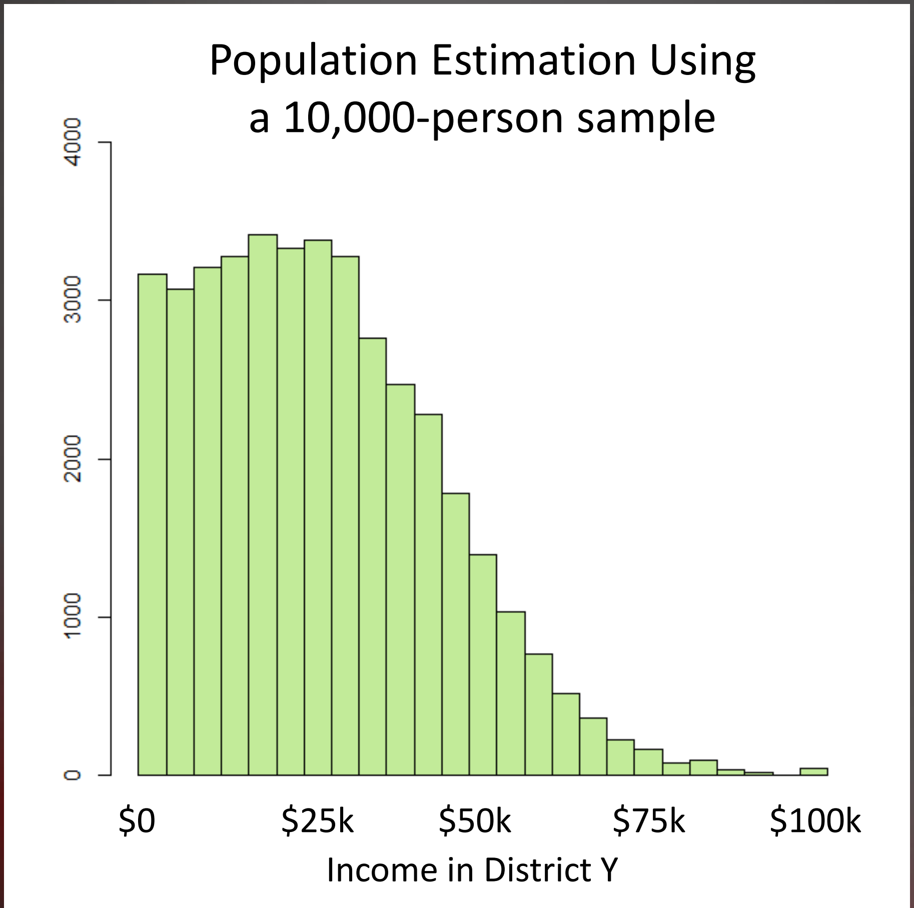
One of Gertrude's colleagues has some more time than Gertrude did, and independently surveys 2,000 people. She uses the properties of this 2,000 person sample to estimate population properties.

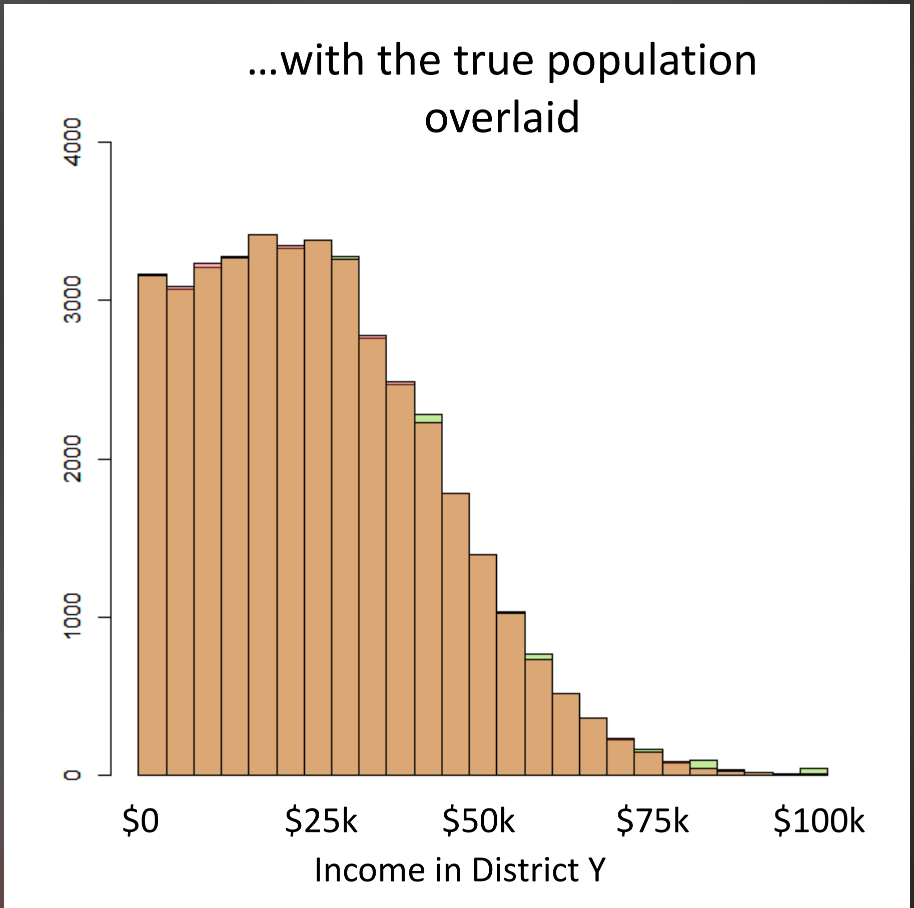




Sampling error in a histogram

Lastly, another colleague surveys 10,000 people, and the estimate from this sample proves to be the most accurate.



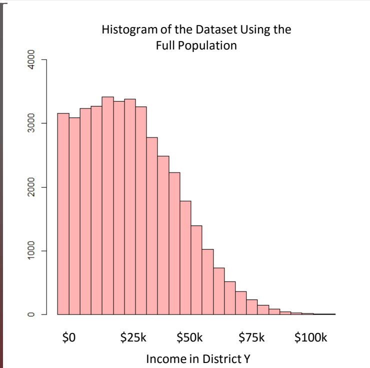


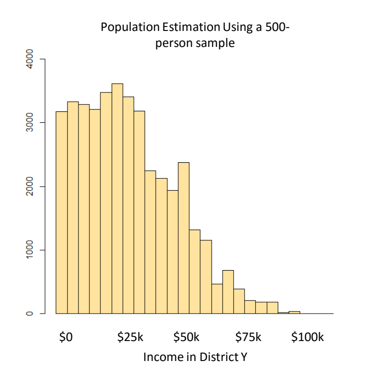
Sampling error in a histogram

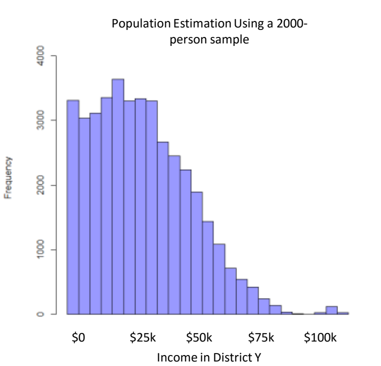
The chart below shows the different means we get from each representation of the population.

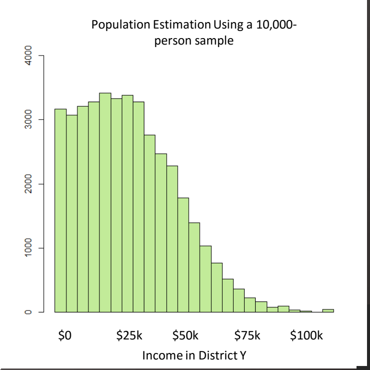
Researcher		**Gertrude**	**Colleague 1**	**Colleague 2**
Sample Size	**(40,000)**	**500**	**2000**	**10,000**
Average in USD	26,000	23,134	27,487	25,567
Difference*	(0)	-2866	+1487	-433

*the empirical mean found by each researcher minus the true mean, 50,000.









Sampling error in a histogram

<u>Keep in mind that so far we've only taken one sample at each sample size, and that there are many other possible random samples.</u>

The chart below shows some other possible averages that each researcher could've found through the same methods used so far.

Researcher		**Gertrude**	**Colleague 1**	**Colleague 2**
Sample Size	**(**40,000**)**	**500**	**2000**	**10,000**
Average in USD	26,000	23,134	27,487	25,567
Error	(0)	-2866	+1487	-433
Average in USD	26,000	29,965	28,201	26,019
Error	(0)	+3965	+2201	+19
Average in USD	26,000	27,178	27,203	25,887
Error	(0)	+1178	+1203	-113

Sampling error in a histogram

Despite these errors, these averages would be suitable for conventional use in research and analysis, and offer nearly the same information as if they'd all returned the true population mean.

We know this because of the significance test, a very common practice. That measurement is calculated using the standard error of the sample, and determining if the returned mean is 'too far' from the mean we expect ($50k).

95% confidence

Social scientists and statisticians are often satisfied with a 95% confidence level in measurements.

“Confidence level” refers to the statistically-computed likelihood that a sample statistic accurately represents the population statistic it estimates.

Based on that measure, and the standard deviation of Gertrude’s team’s samples, none of the results are significantly different from their respective population mean, and would all be treated as essentially the same answer.

Section 2.3

SAMPLING ERROR AND INFERENCE IN SOCIAL SCIENCE

Means Histograms CDFs

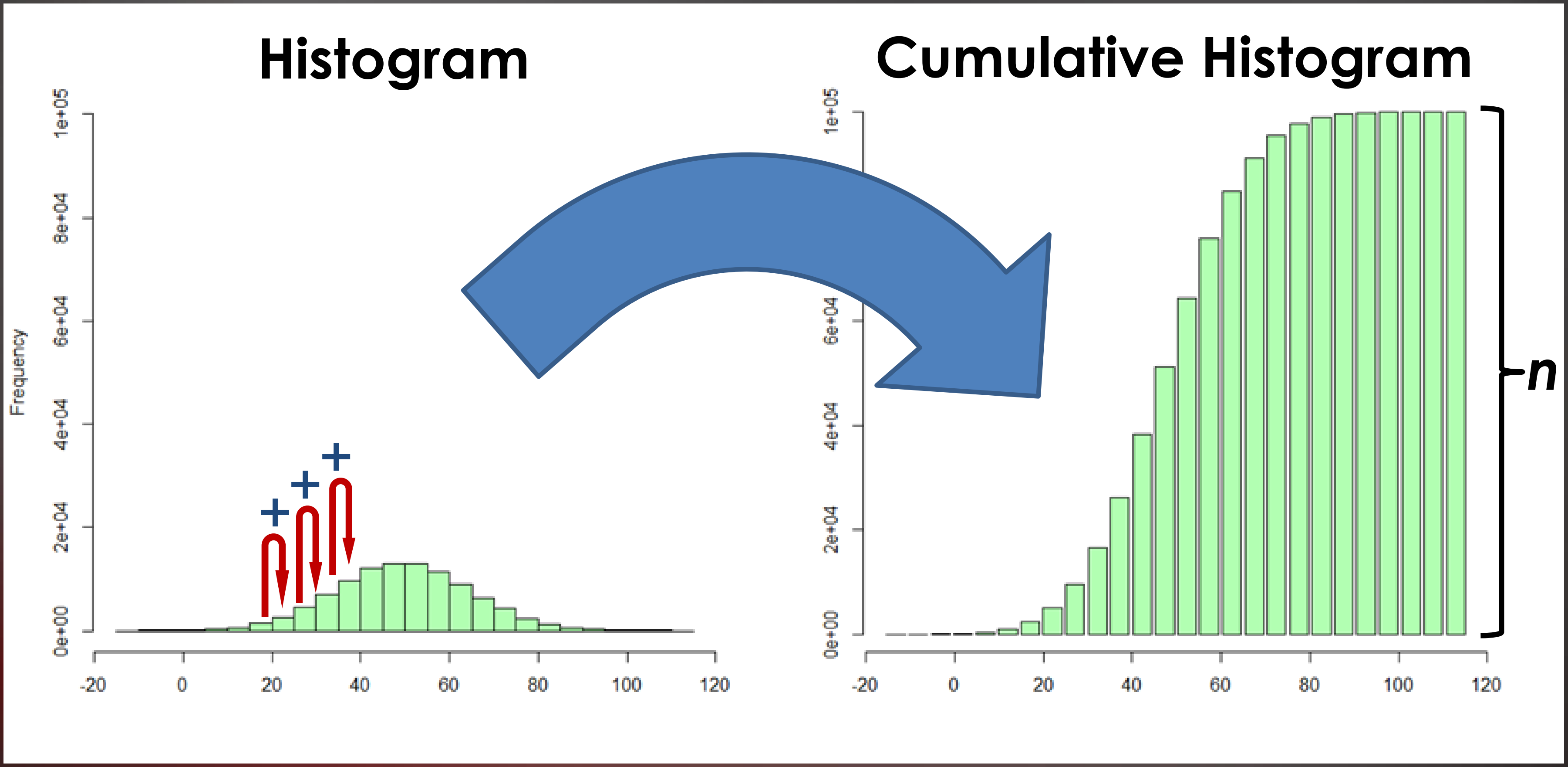
Sampling error in CDFs

Thus far, we've covered the behavior of sampling error in histograms and the averages they return. Something slightly more complicated is the behavior of sampling error in accumulation, as it is added together.

This situation occurs in cumulative histograms and their more common counterpart, cumulative density functions, or CDFs. CDFs are used mainly to compute the medians and other quantiles of datasets. In the following slides we'll explain the construction of CDFs with an eye on how the randomness of sampling error affects their accuracy and utility.

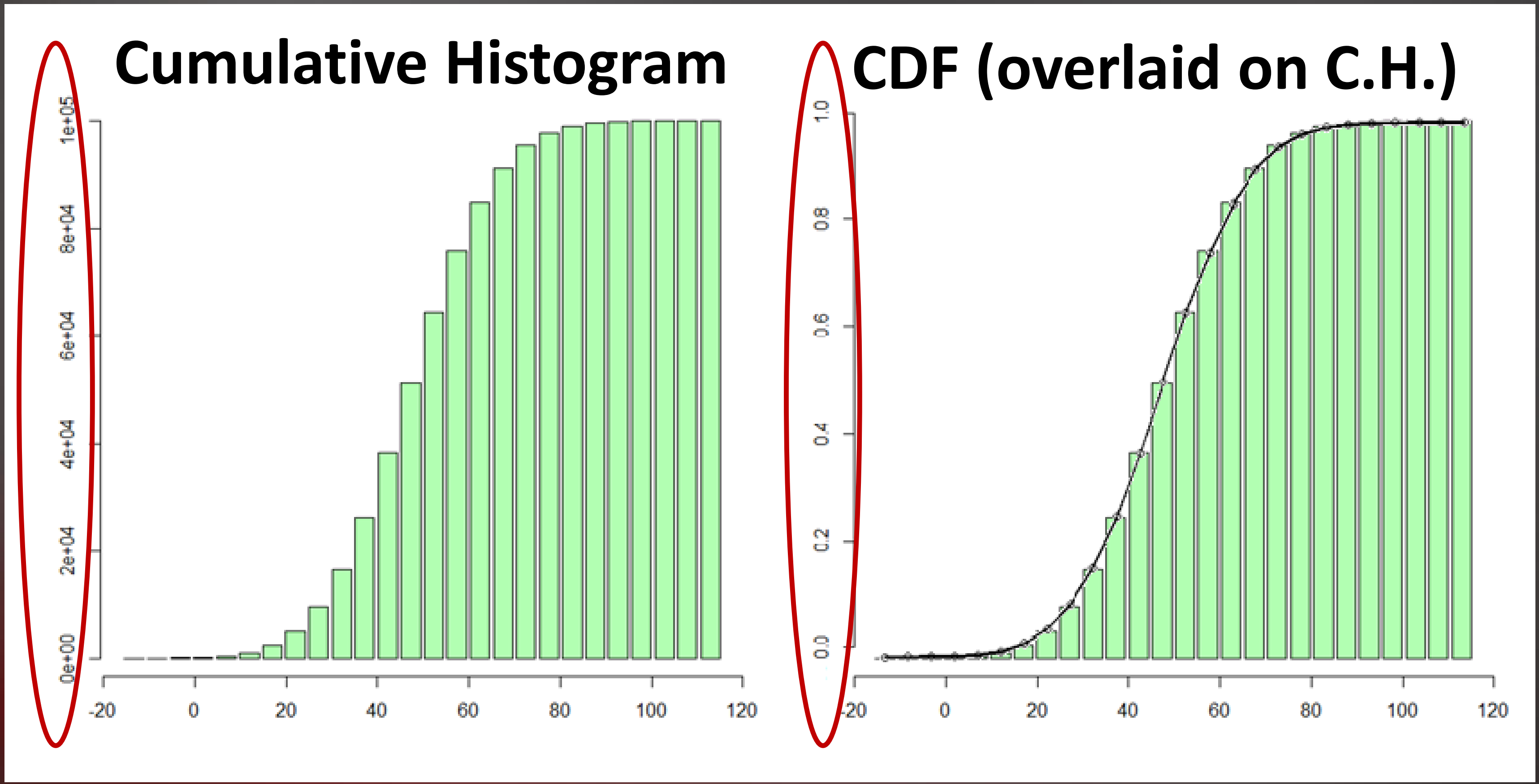
Cumulative Histograms

First, sequentially adding the bars of a histogram gets us a cumulative histogram. The final bin of a cumulative histogram will always equal the size of the original dataset.



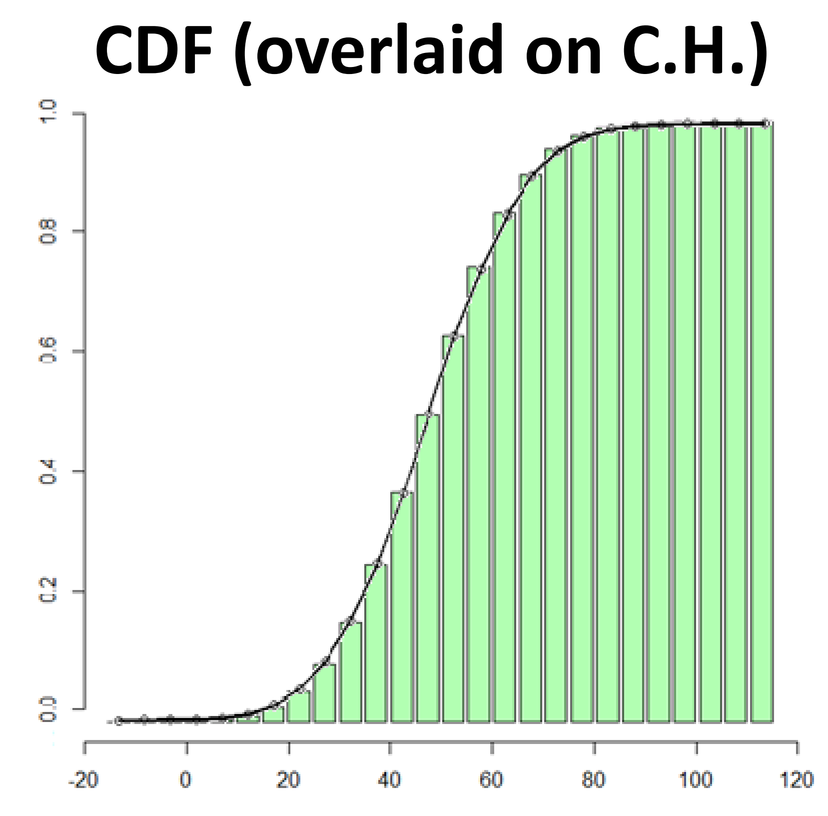
Cumulative Density Functions

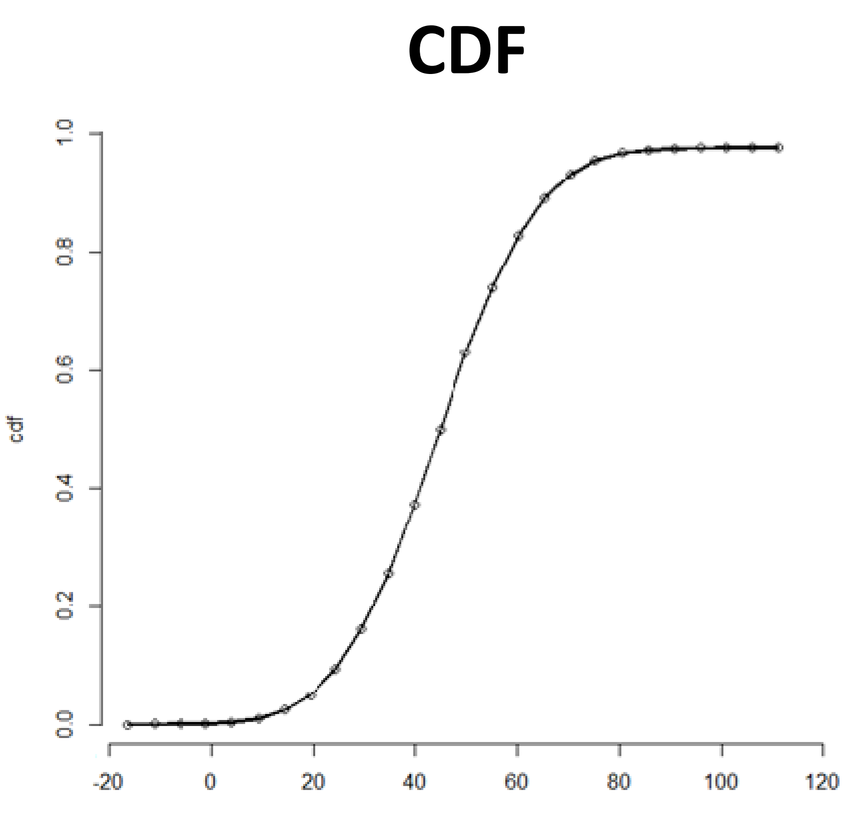
Dividing the cumulative histogram's bin-counts by the total number of data points, we normalize to [0,1] and approximate a cumulative density function (CDF).



Cumulative Density Functions

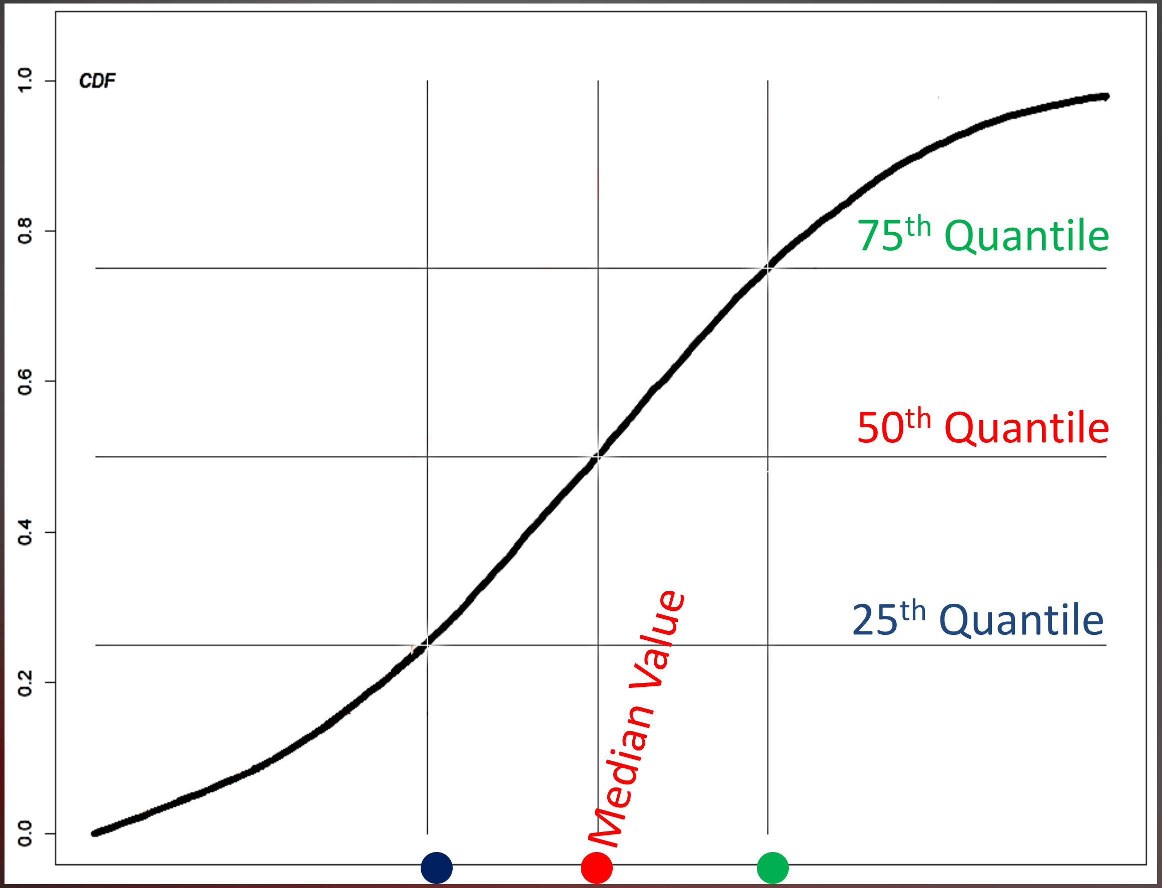
Removing the underlying bars used to compute CDFs, we see the more commonly-used curve-form of a CDF. We'll continue with curves like this.





Cumulative Density Functions

Formally, CDFs are integrations of a random variable's probability density function*. Their main use is in how they represent proportions of datasets.



Thanks to normalization, they can be easily used to find quantiles/percentiles of a random variable's values, as seen here.

*Note on Empirical CDFs

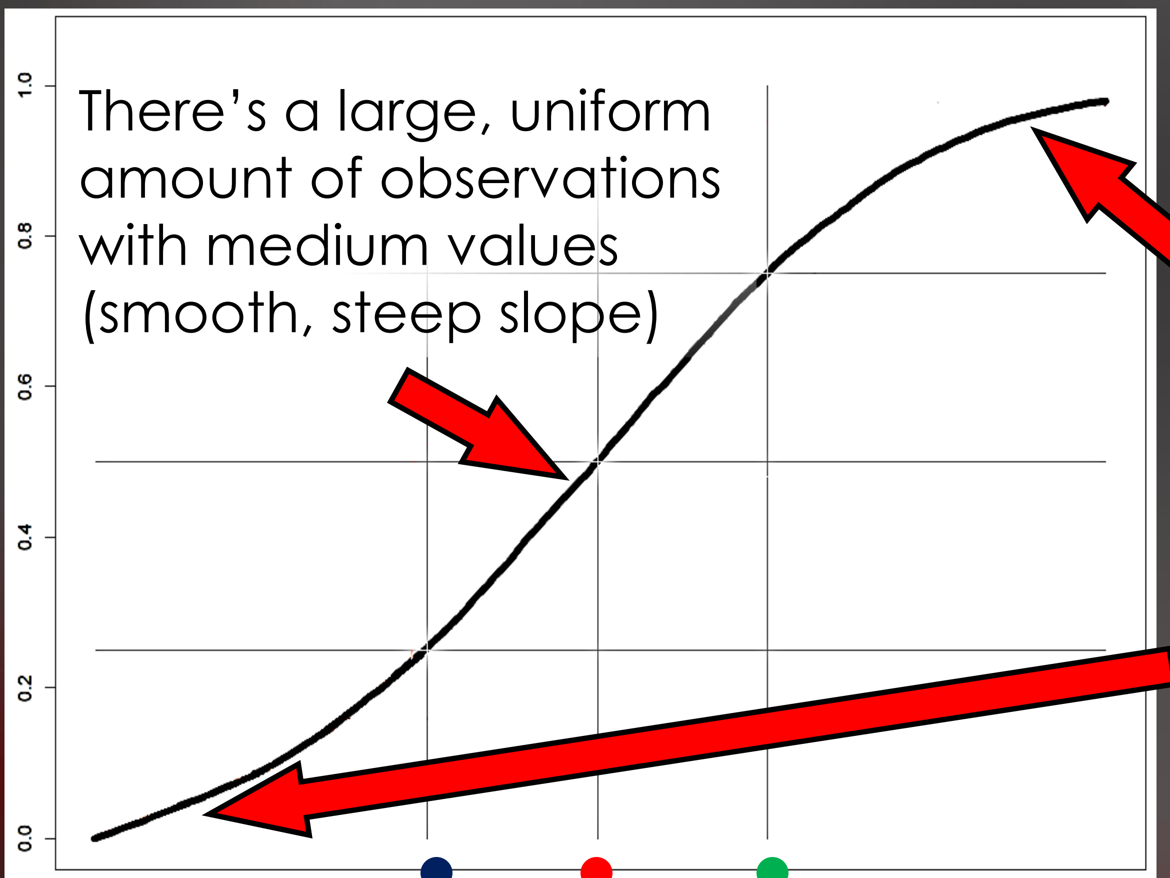
Here, we note a technical point.
There is a difference between CDFs generated from sample histograms and CDFs generated strictly as integrations of PDFs, the latter of which is truly a CDF.

That is, a true CDF is the underlying *function* dictating how the graphical curve appears, whereas an empirically-derived "CDF" created from a histogram is actually a set of points following a pattern dictated by an unknown underlying function.

For most applied empirical usage, social scientists are actually using empirically derived CDFs, and we continue that tradition in work on applied differential privacy. That is, we are not considering underlying mathematical functions.

Cumulative Density Functions

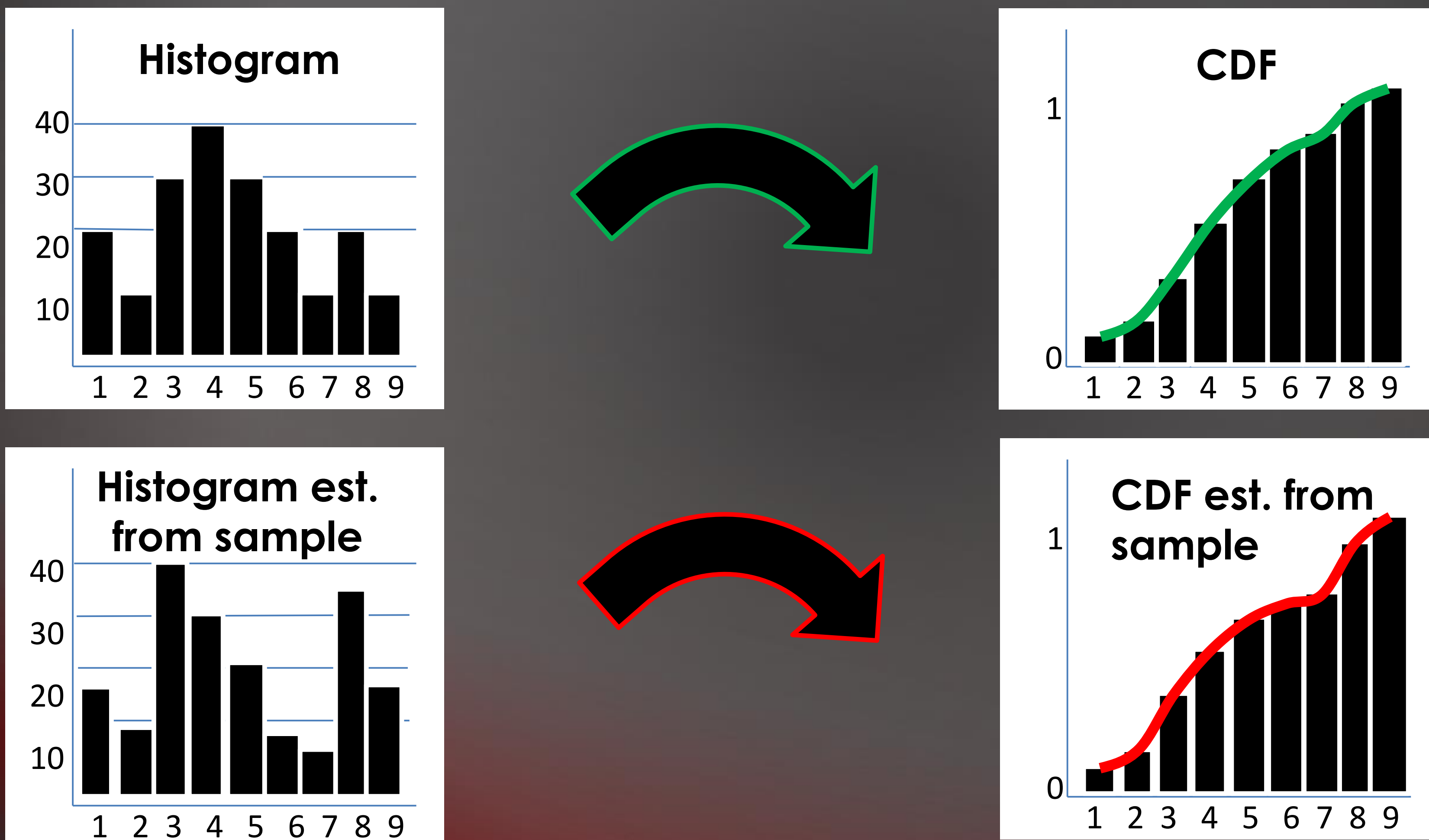
As a final point, CDFs can show general trends in data as well. For example, we see that this distribution is roughly Gaussian.



There are fewer observations with high and low values (small slope).

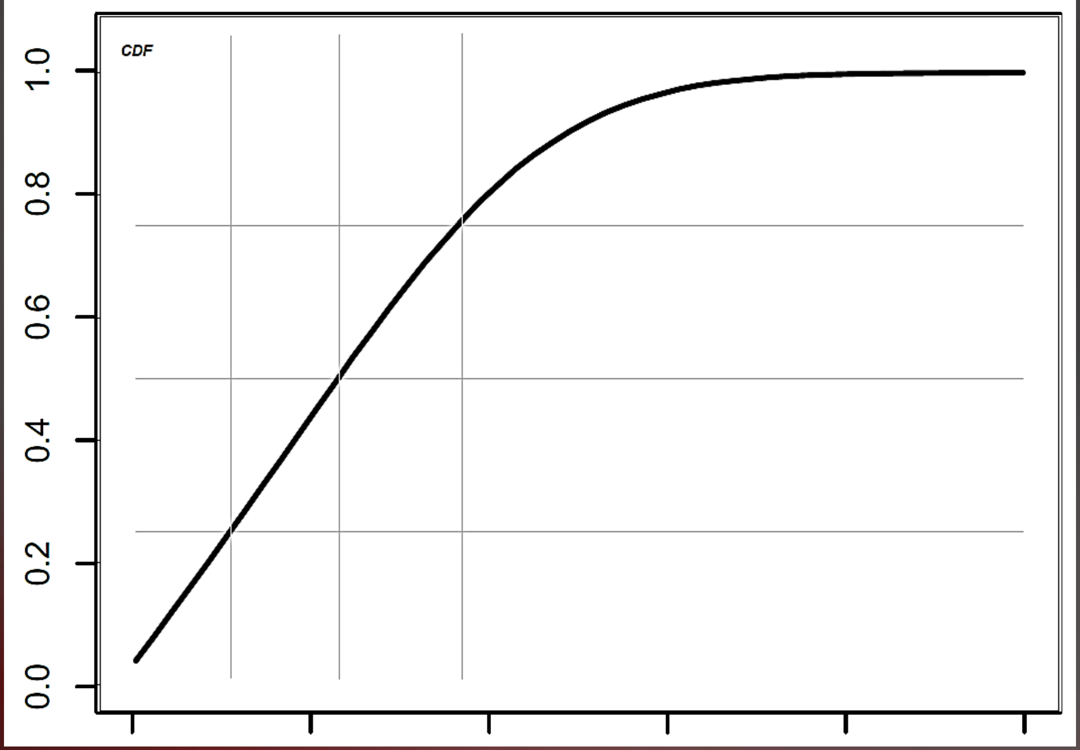
Sampling Error in CDFs

CDFs contain the same amount of sampling error as the histograms and PDFs that they are created from.



Sampling error in a CDF

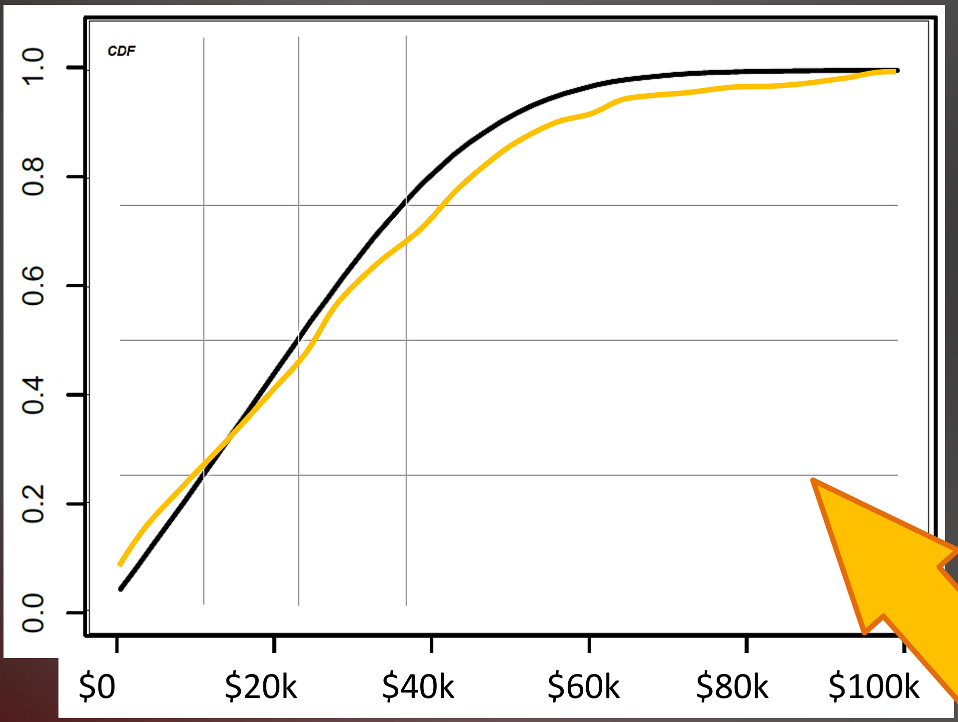
In the following example, we'll present a few CDFs constructed with random sampling noise. For reference, we'll sometimes overlay these errored-CDFs on top of a their true CDF (of the entire population), which will most often take the form below.



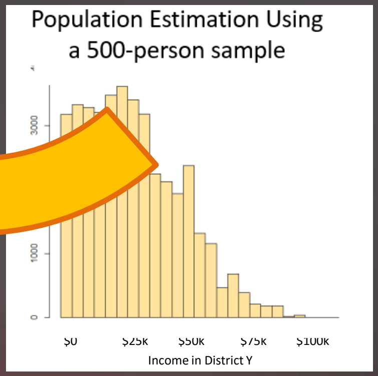
In each image, the horizontal gridlines represent the 25th, 50th, and 75th percentiles, and the corresponding vertical gridlines trace the corresponding values in the true dataset. This helps to visualize manifest error.

Sampling error in a CDF

Gertrude, our researcher from earlier, needs to find the *median* income of District Y, and expects she'll need to find other quantiles later.

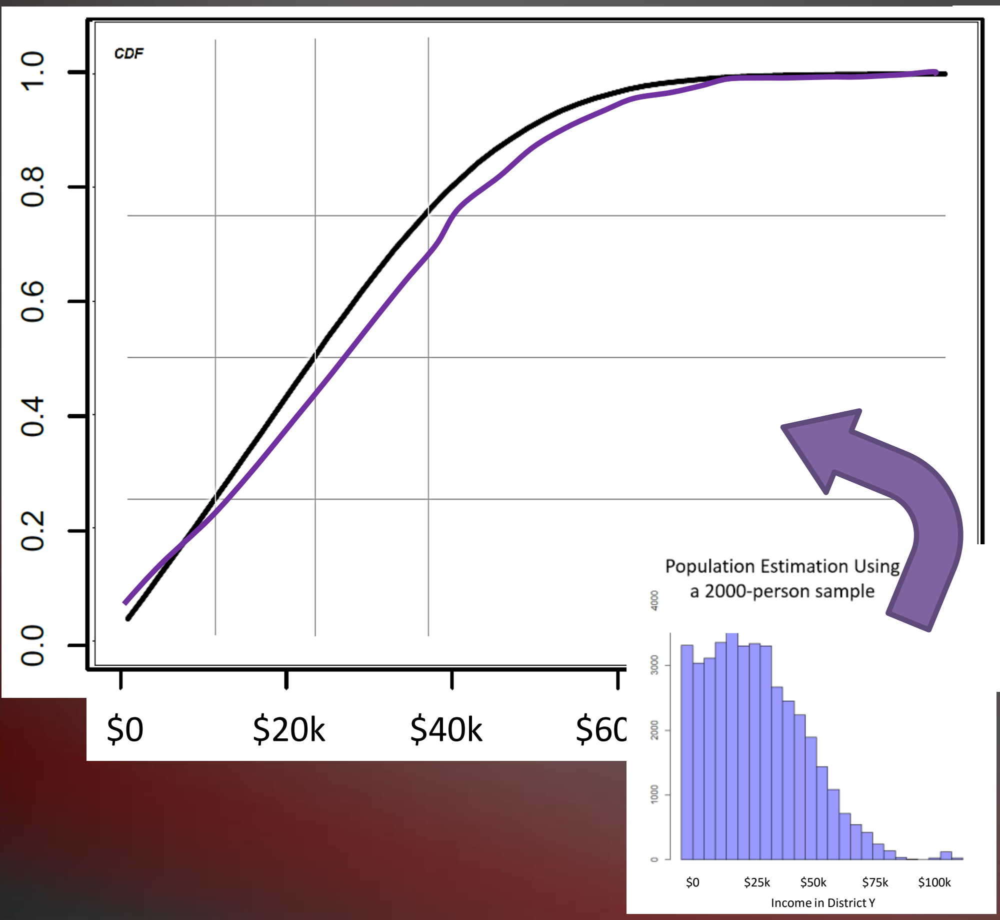


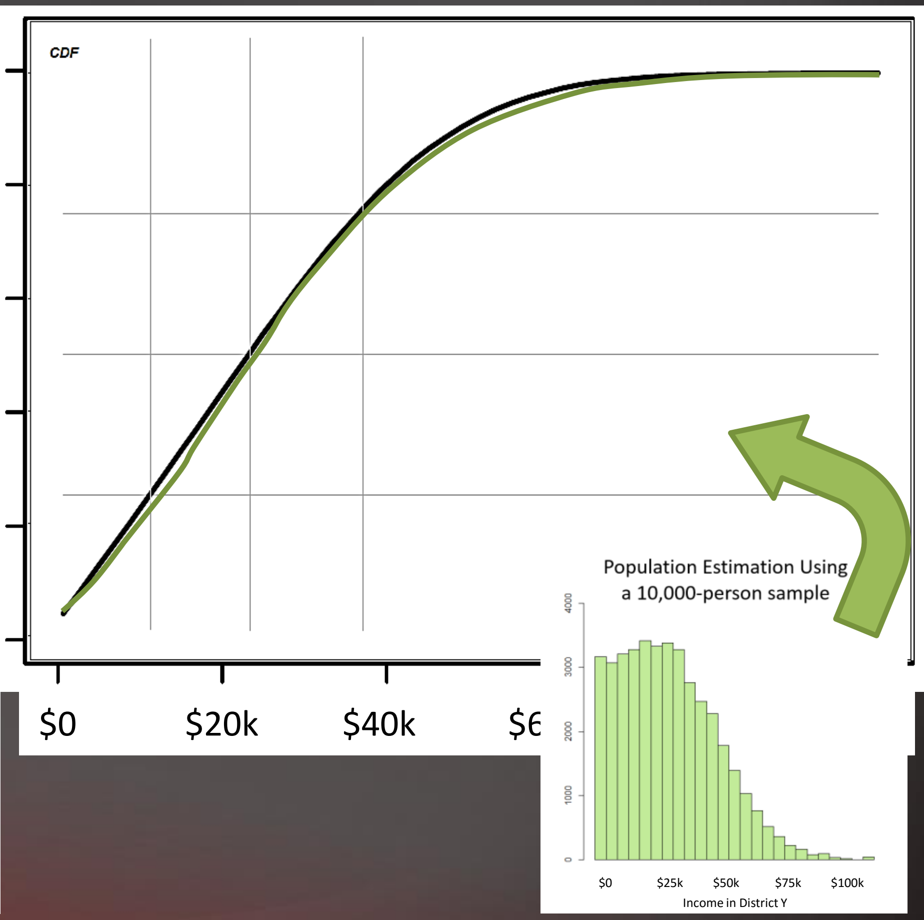
Using the histogram she made earlier from her 500 person survey in District Y, Gertrude uses statistical software to build a CDF. Her CDF is an estimation of the true CDF of District Y's entire income distribution.



Sampling error in a CDF

Gertrude's two colleagues want to check her answer. They each use the histograms constructed earlier to make a CDF and find estimate the median income of district Y.



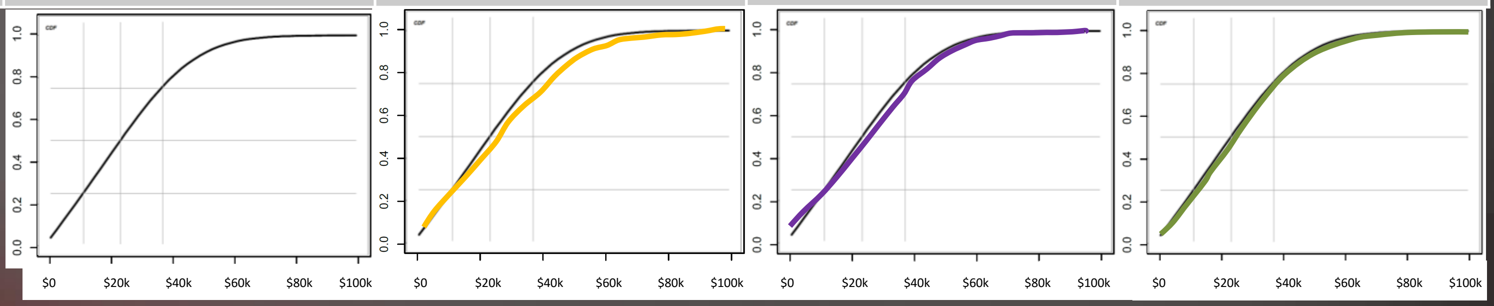


Sampling error in a CDF

The chart below shows the different medians we get from each sample.

Researcher		**Gertrude**	**Colleague 1**	**Colleague 2**
Sample Size	**(40,000)**	**500**	**2000**	**10,000**
Median in USD	23,500	26,000	24,700	23,800
Difference*	(0)	+2500	+1200	+300

*the empirical median found by each researcher minus the true mean, 23,500.



Sampling error in a CDF

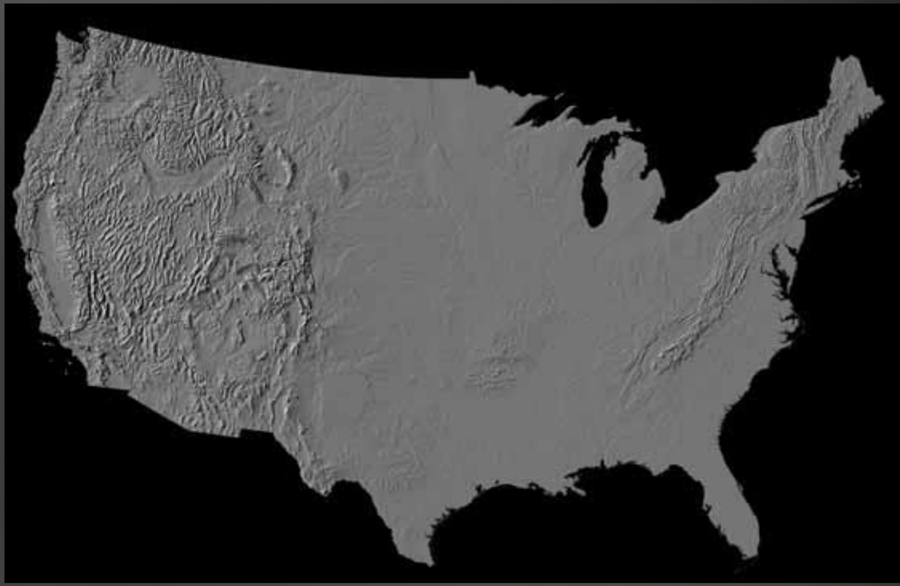
Keep in mind that so far we've only taken one sample at each sample size, and that there are countless other possible random samples.

The chart below shows some other possible medians that each researcher could've found through the same methods used so far, with differences being due to sampling error.

	Researcher		**Gertrude**	**Colleague 1**	**Colleague 2**
	Sample Size	**(40,000)**	**500**	**2000**	**10,000**
Sample 1	Median in USD	23,500	26,000	22,300	23,900
	Error	(0)	+2500	-1200	+400
Sample 2	Median in USD	23,500	22,300	24,100	23,000
	Error	(0)	-1200	+600	-500
Sample 3	Median in USD	23,500	25,100	24,600	23,200
	Error	(0)	+1600	+1100	-300

Managing Random Noise in Modern Social Science

As a final note, it's worth noting that the United States Census typically only surveys less than one percent of the U.S. population. This is what our arguably strongest socio-economic data comes from.



With that, we hope that randomness and random noise are recognized as an immutable and constant part of statistical work. We now move on to explaining ***differential privacy***, a method of using randomness and uncertainty to the advantage of privacy.

Section 3

DIFFERENTIAL PRIVACY: INTUITION

Why Data Privacy?

Before explaining what differential privacy does, it's important to understand the motivation behind it.

First, we must see that there is a vast amount of personally identifiable and *sensitive* information online. Things like names, addresses, medical records, and financial information are collected by firms, banks, hospitals, schools, and even the government.

Why Data Privacy?

Data's Value

Large datasets that include sensitive information can be highly valuable to researchers. Consider an epidemiologist seeking trends in citizens' health, or an economist studying the ownership of volatile financial assets. We'll continue to think about these "good guys", the researchers.

Accessing Data

Researchers who want to access sensitive data must go through long, time-consuming, and potentially expensive processes to confirm that the data will be safe in their hands. Some data is simply inaccessible to even the best-intentioned researchers.

Why Data Privacy?

Data's Danger (to the public)

For reasons of profit, personal gain, or general malign motivation, criminals or predatory organizations can use this personal data as a tool. These individuals and organizations are often referred to as "adversaries."

For example, medical information can be used by adversaries for blackmail, harassment or persecution. Financial data can be used by adversarial firms for discriminatory pricing or advertisement.

Why Data Privacy?

Data's Danger (to the researcher)

At the same time, these risks can make data a liability for the researchers themselves, and their institutions. Holding and sharing sensitive data opens researchers to the risks associated with failing to uphold related legal and ethical requirements.

Poorly handled sensitive information can also jeopardize the reputation of researchers, institutions, and the research community at large.

Why Data Privacy?

What is wrong with current data privacy standards?

At present, standards for protecting the privacy of data are often weak. For example, removing names and addresses from datasets may at first seem sufficient to protect individuals, but this has been disproven by powerful examples.

The same can be said of many other typical privacy-preservation techniques. For more, see related publications here.

Why Differential Privacy?

Conflation of privacy goals and methods

Part of the weakness in typical privacy-preservation techniques comes from conflating the _goal_ of privacy with the _methods_ of privacy preservation.

For example, a researcher removing names from a dataset may think that the removal of names itself is the privacy goal, when reality, it's the *method*. The *goal* in this case might be that no individual in the dataset can be re-identified, or that no one would experience harm from the dataset.

Why Differential Privacy?

Separation of privacy goals and methods

Differential privacy is founded on a purely mathematical privacy goal, and methods are developed to meet that standard.

For more on the mathematical foundation of differential privacy, you can find related documents and explanations here.

What is Differential Privacy?

The Methodology of Differential Privacy

At the most basic level, differential privacy refers to methods that introduce *random noise* into statistical analyses.

Differential privacy requires that for an given dataset, if a single person's data is added to or removed from that dataset, statistics computed from that dataset with differential privacy should be statistically indistinguishable before and after that change.

Furthermore, the combination of several statistical analyses that each satisfy the requirement of differentially privacy results in a (compound) analysis that satisfied differential privacy (albeit, with weaker guarantees). This is known as *composition*.

Why Differential Privacy?

Accessing Sensitive Data with Differential Privacy

Differential privacy can help remove the barrier between researchers and sensitive data while providing a strong protection for the privacy of individual data contributors.

Why Differential Privacy?

Accessing Sensitive Data with Differential Privacy

By properly utilizing differential privacy, researchers are able to investigate sensitive datasets *before* going through the long process of seeking full access.

Through the platform we develop, researchers may ask for statistical information regarding a sensitive dataset, such as means, histograms, or regressions. With differential privacy, the resulting statistics are slightly obscured through a mathematically precise method.

These results can be highly valuable to researchers deciding whether or not to access a dataset, but negligibly useful to adversaries seeking personal information.

What is Differential Privacy?

Other documents from this group provide mathematical definitions for the inner-workings of differential privacy. This document will simply cover the intuition of differential privacy as far as is useful for statistical work.

Slightly more in-depth explanations can be found in the appendix at the end of this document.

Section 4.1

DIFFERENTIAL PRIVACY: APPLICATION

Means | Histograms | CDFs

Using Differential Privacy

We'll now offer a simple example. We return to Neil, a social scientist previously concerned with sampling error. Remember that Neil was seeking to find the mean number of social groups that individuals in District X identify with.

Number of Social Groups each Person Identifies with in District X

3	5	3	5	3	3	4	5	4	5	3	5	5	5	3	0	8	6	8	6	9	6	4	0	4	7	6	8	7	2
1	0	5	5	3	3	5	7	0	3	2	3	1	3	0	1	5	7	2	4	6	9	7	0	9	9	4	5	9	1
4	4	2	6	7	9	3	0	9	3	5	6	3	2	9	5	8	6	7	5	3	4	3	9	5	0	3	3	2	0

Above is the full population of District X (mean 4.4)

															0								0						2
1	0								3			1							4						9				
		2					0						3				6			3				5	0			2	0

And here is Neil's random sample (mean 2.3)

Recap: Sampling error

To reiterate, here are the random samples of Neil's three colleagues, and their respective sample means.

Number of Social Groups each Person Identifies with in District X

3							5						5							9	6				7			7	
							7				3					5					9				9				
	4																	7		3			9	5	0				

Sample mean: 5.72

					3					3			5					8									8		
							7				3					5							0	9					
							0							9	5					3	4				0	3	3		

Sample mean: 4.33

															0							4				6			2
1		5						0	3									2			9		0						
				7						5				9						3					0			2	0

Sample mean: 3.22

Using Differential Privacy

Now, suppose Neil learns that a university has already conducted an identical survey in a nearby area, District Z.

Neil believes that District X and District Z have similar populations, so he wants to access the university's data for his research.

However, the university labels that dataset *sensitive*, and will not release the data without a year-long legal process to verify Neil's credentials.

Neil wants to make sure that going through the process is worthwhile. If District Z is actually quite different from District X, he isn't interested. He decides to investigate the data using differential privacy.

Using Differential Privacy

Below, we see the university's data. This is the data that Neil <u>does not</u> currently have direct access to, due to legal and privacy concerns.

Number of Social Groups each Person Identifies with in District Z

3	5	3	2	3	3	5	5	4	5	1	2	3	3	6	5	8	0	7	3	6	4	4	2	4	5
1	0	5	5	3	3	5	7	0	3	2	2	4	3	1	5	4	1	3	3	1	9	7	0	9	9
4	4	6	1	7	9	3	0	9	3	5	6	3	2	9	5	8	6	5	5	2	1	1	9	1	3

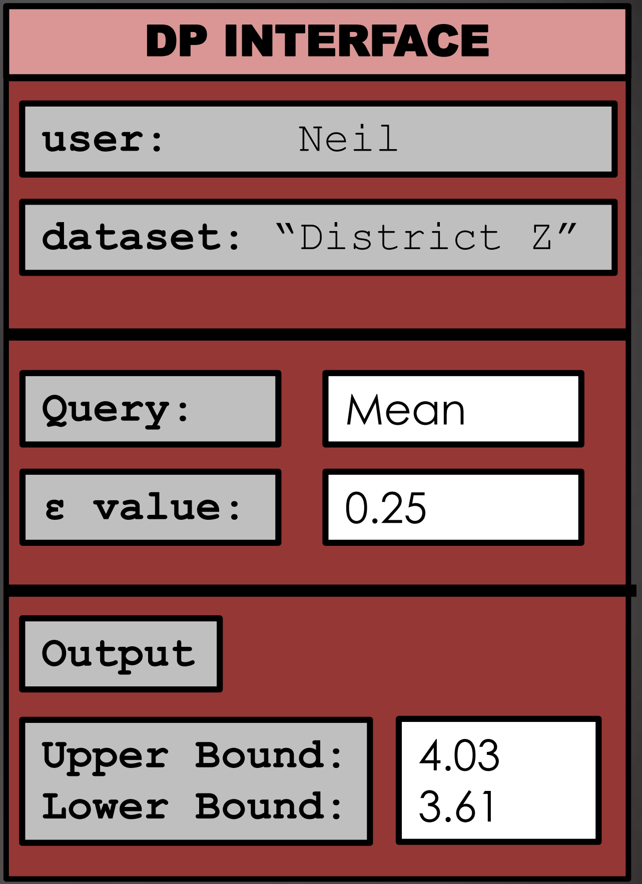
Since we see the exact data, we can compute the exact mean: 4.01

Using Differential Privacy

The university directs Neil to their differential privacy platform.

Neil does not see the data directly. He specifies two things: the statistic he is interested in, and the ε value, which we'll return to.

He receives an output of "likely between 3.61 and 4.03", which is a *differentially private approximation* of the mean. So how did we arrive at this?



Using Differential Privacy

Behind the scenes, Neil's query was sent through the "DP Interface" to the true data.

Number of Social Groups each Person Identifies with in District Z

3	5	3	2	3	3	5	5	4	5	1	2	3	3	6	5	8	0	7	3	6	4	4	2	4	5
1	0	5	5	3	3	5	7	0	3	2	2	4	3	1	5	4	1	3	3	1	9	7	0	9	9
4	4	6	1	7	9	3	0	9	3	5	6	3	2	9	5	8	6	5	5	2	1	1	9	1	3

The interface calculated its answer by computing the mean and then adding proportional random noise. That is:

$$\text{DP mean} = \frac{\sum(xi)}{n} + \text{noise}(\boldsymbol{\varepsilon})$$

$$= \frac{313}{78} + \text{noise}(\mathbf{0.25})$$

$$= 4.01 + (\text{-}0.19) = 3.82$$

Using Differential Privacy

$$\text{DP mean} = \frac{\sum(xi)}{n} + \text{noise}(\boldsymbol{\varepsilon})$$

$$= \frac{313}{78} + \text{noise}(\mathbf{0.25})$$

$$= 4.01 + (-0.19) = 3.82$$

$$3.82 \;-\; 0.21 = 3.61$$

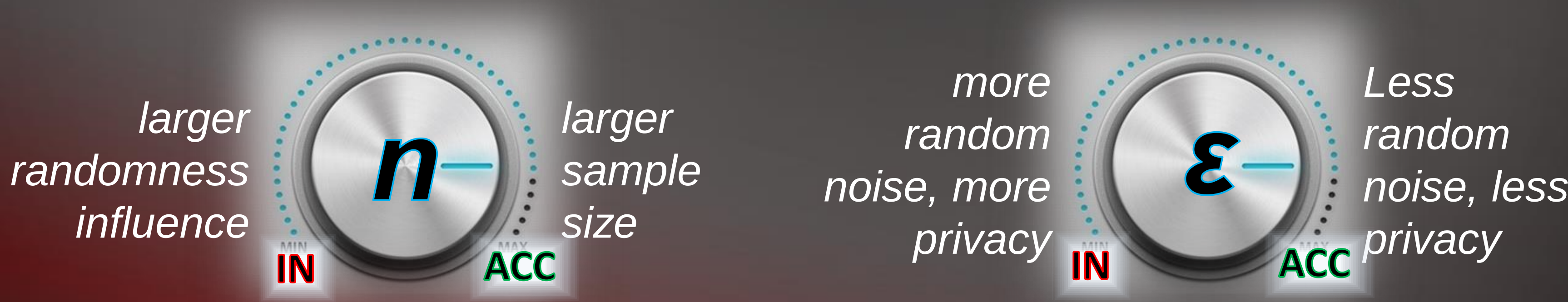
$$+\; 0.21 = 4.03$$

This initial estimate of the mean, 3.82, is not very informative without its confidence intervals, or range or possible values. Using 95% confidence bounds based the ε parameter of differential privacy, the interface adds and subtract 0.21 from this answer, giving us out upper and lower bounds of 3.61 and 4.03

ε, the "Knob" of DP

Users like Neil have some control over the level of random noise introduced by differentially private computations. As mentioned earlier, we can think of ε (pronounced "epsilon") as a knob we can turn to tune our differentially private results.

Just as we can turn the "knob of sample size" up or down to increase or decrease the precision of our statistics, <u>we can increase or decrease the scale of DP random noise by increasing or decreasing ε</u>. Further, they can be turned simultaneously.



ε, and request limits

In the next slides, we'll look at what *could have* happened if Neil had chosen different levels of ε.

It's important to understand that the following examples are only hypothetical.

Neil, like all DP-users, cannot request the DP-statistic from a dataset more than once. Multiple requests could compromise the privacy of the data.

ε, the "Knob" of DP

Here is the mean from above (*4.01*) approximated with different levels of ε. With lower ε (left), we're more likely to receive larger noise levels. <u>This is only for educational purposes; Neil can only see one answer</u>.

DP INTERFACE	DP INTERFACE	DP INTERFACE
user: $\text{Neil}_{\text{hypothetical}}$	**user:** Neil	**user:** $\text{Neil}_{\text{hypothetical}}$
dataset: District Z	**dataset:** District Z	**dataset:** District Z
Query: Mean	**Query:** Mean	**Query:** Mean
ε: 0.1	**ε:** 0.25	**ε:** 0.5
Output:	**Output:**	**Output:**
Upper Bound: 4.85	**Upper Bound:** 4.03	**Upper Bound:** 4.09
Lower Bound: 4.15	**Lower Bound:** 3.61	**Lower Bound:** 3.93

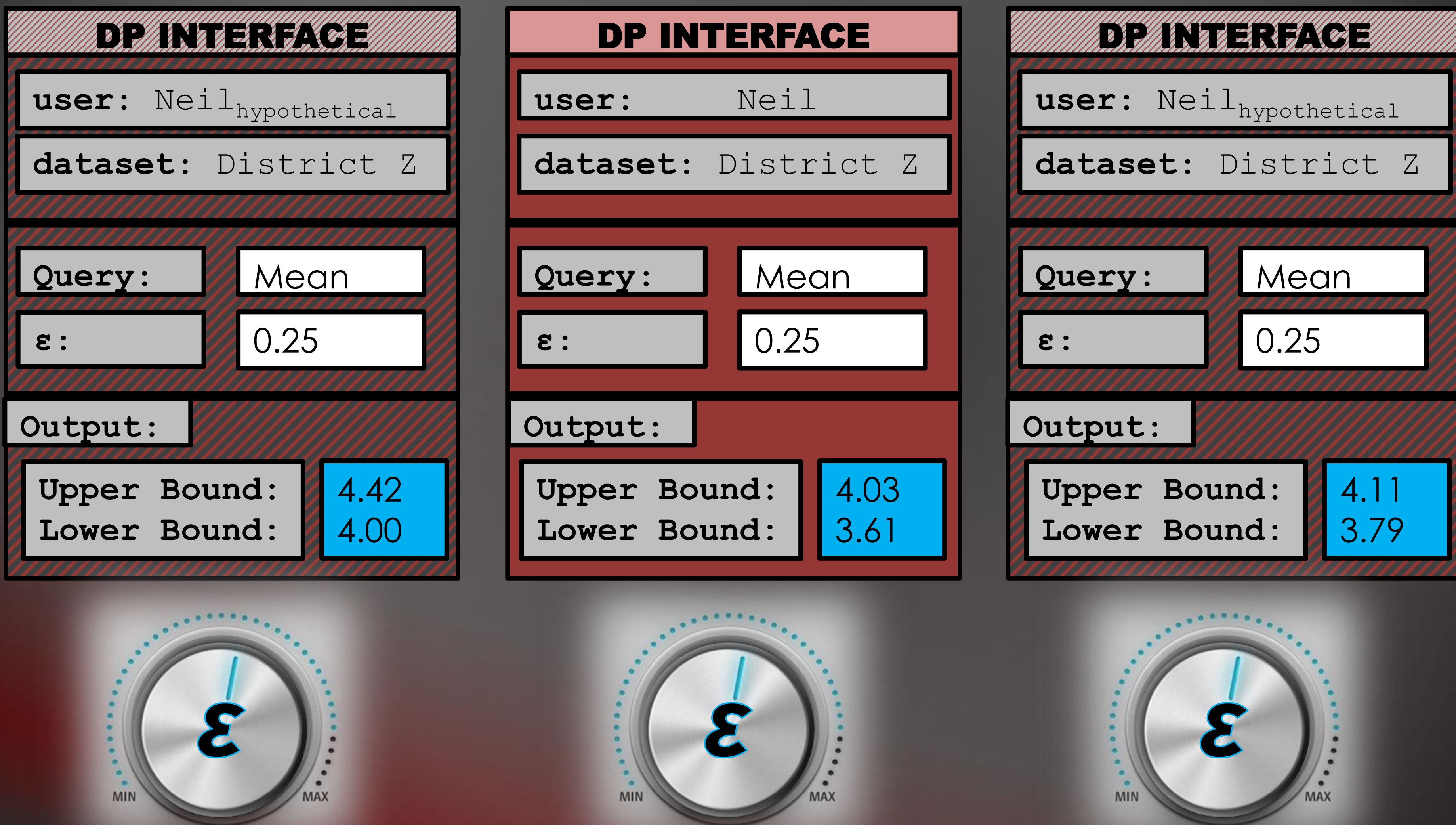
Randomness in Differential Privacy

This leads us to another key point. Since the methodology of differential privacy relies on random noise addition, differentially private approximations can produce various results, even under identical conditions.

On the next slide, we'll examine this property of differential privacy.

ε, the “Knob” of DP

First, Neil knows that the added noise is *random*, repeating the same computation (even if it’s with the exact same value of ε) can produce different answers. These answers would have the same size confidence interval.



ε, the "Knob" of DP: Why?

At this point, a user may wonder why anyone would use a low ε value, since higher epsilon values generally get more useful answers.

District Y	Actual	Neil's Result	Hypothetical Answers (from previous slides)			
ε value	ε = N/A	ε = 0.25	ε = 0.1	ε = 0.25	ε = 0.25	ε = 0.5
Mean value	**4.01**	3.82	4.45	4.21	3.90	4.05

The reason is that differential privacy systems have a limit on how high you can set ε. This limit includes the cumulative ε used across computations. Thus, we often refer to a "privacy budget" that corresponds to the maximum "accumulated **ε**" allowed for statistical queries into a dataset or DP-interface. Users want to conserve some of their ε budget for future use, and thus might want to use low ε values.

ε, and "alpha" choices

As a side note, many Differential Privacy interfaces automate the selection of ε. Instead of asking users to enter an ε value, they might ask users for "alpha" value. Alpha represents the accuracy of an outputted DP statistic. When a user inputs an alpha value, the system automatically uses it to select the epsilon value required.

However, ε is at the root of all DP calculations, and alpha is merely a tool used to help determine its value. Therefore, we'll continue using ε in our examples.

Differential Privacy: means

Returning to Neil's situation, Neil has now seen his differentially private approximation of the mean, "most likely between 3.61 and 4.03".

From that, he safely assumes that the true mean number of social groups that residents of District Z belong to is similar to that of District X. He decides that this data would be helpful to his research, and he decides to file for full access to the university's dataset.

Differential Privacy: means

Later, Neil learns of District W, a third nearby area that the university has information about. Neil asks for one differentially private mean for District W. He sets ε = 0.25, and his answer is 5.21. Neil knows that the true answer is probably roughly between 4 and 6, and so he asks for District W's data as well.

Lastly, Neil learns of District L, a fourth nearby area that the university has information about. Neil asks for one differentially private mean for District L. He sets ε = 0.25, and his answer is 1.13. Neil knows that the true answer is probably roughly between 0 and 3, and so he does not ask for District L's data.

Section 4.2

DIFFERENTIAL PRIVACY: APPLICATION

Means

Histograms

CDFs

Differential Privacy: Histograms

As we've seen, it was fairly straightforward for Neil to make sense of differentially private means and to incorporate them into his research process.

As we move onto more complex statistics, applying differential privacy becomes slightly more involved.

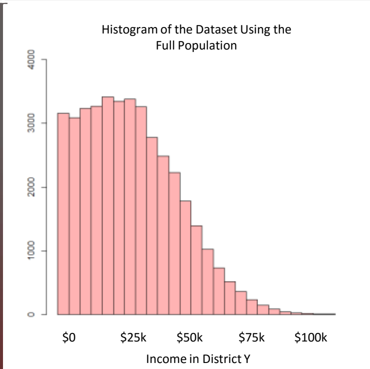
We'll now begin discussing more complex statistics by returning to Gertrude, who makes use of histograms to study income distribution.

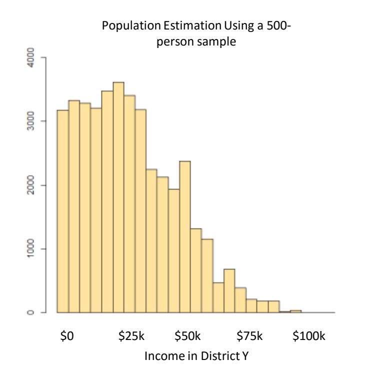
Recap: Sampling error in a histogram

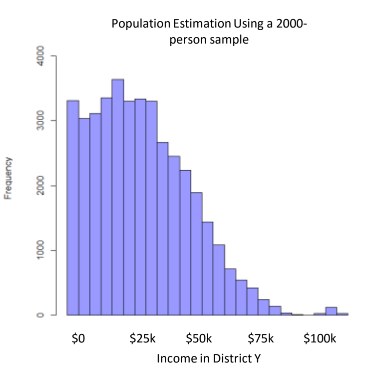
First, recall how sampling error affected the research of Gertrude and her colleagues in District Y. We see the "knob" of sample size in action here, with low sample sizes returning less accurate histograms.

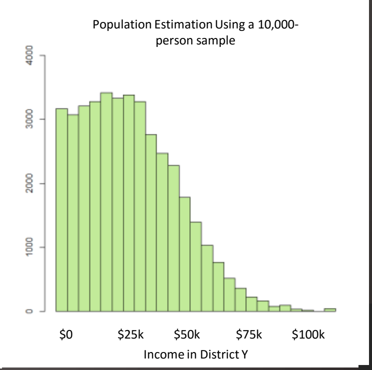
Researcher		**Gertrude**	**Colleague 1**	**Colleague 2**
Sample Size	**(40,000)**	**500**	**2000**	**10,000**
Average in USD	26,000	23,134	27,487	25,567
Difference*	(0)	-2866	+1487	-433

*the empirical mean found by each researcher minus the true mean, 50,000.









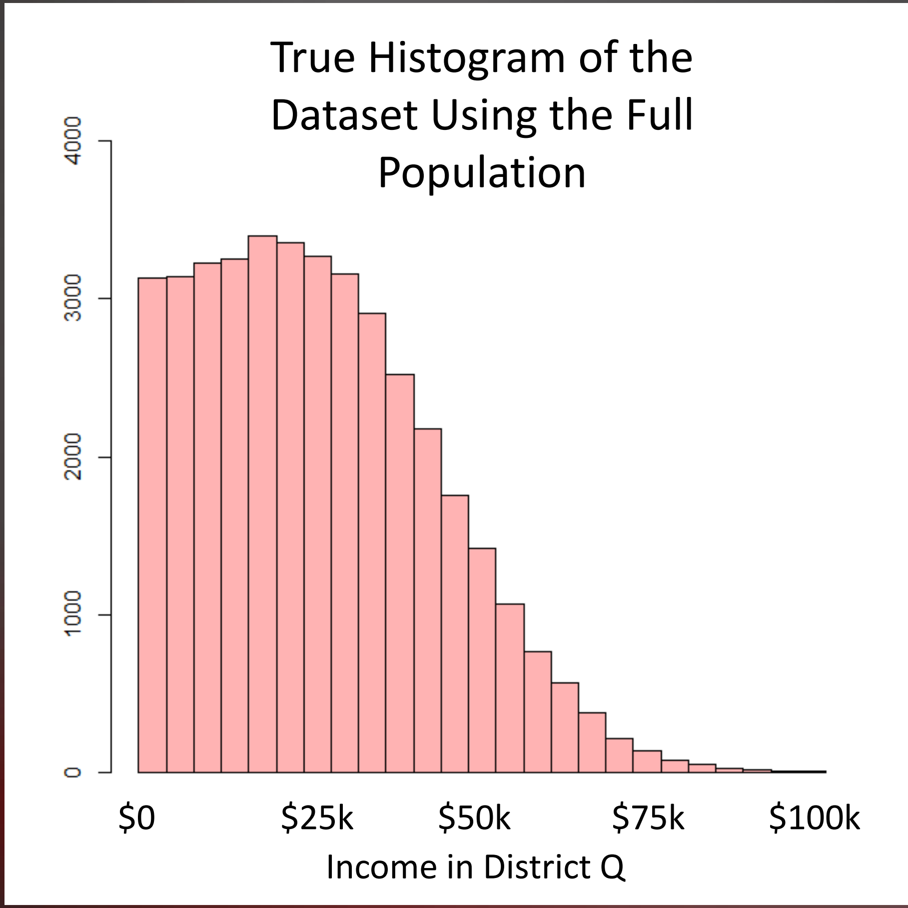
Differential Privacy: Histograms

Like Neil, Gertrude recently learned that a nearby university already conducted a survey on nearly all of the residents of a similar nearby area, District Q. She's curious about their findings, specifically about the income distribution in District Q. If the income distribution of District Q is comparable to that of District Y, then Gertrude is interested in investigating District Q deeply.

The university's reasonable privacy policy prevents them from sharing income information without institutional review, but allows researchers to view statistics, including histograms, with differential privacy. Gertrude decides to peek at District Q's income distribution through a differential privacy interface.

Differential Privacy: Histograms

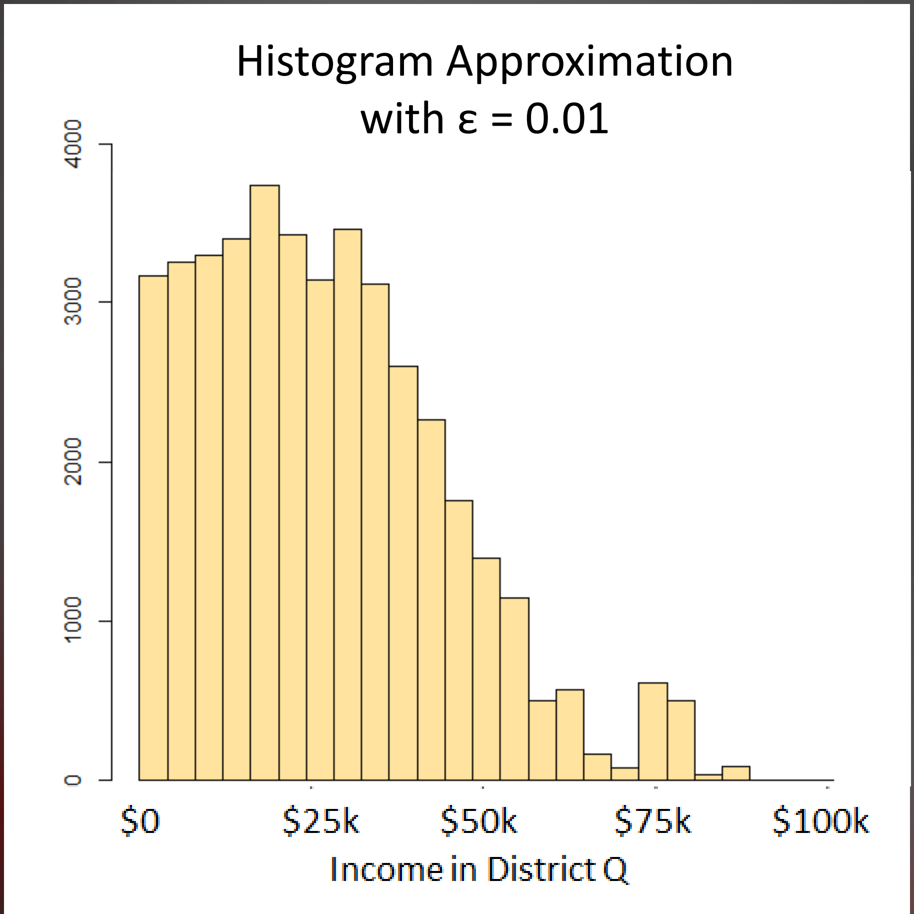
In red is the income distribution of District Q as surveyed by the university. <u>Gertrude doesn't have access to this histogram</u>. Keep this in mind as we follow Gertrude through her research.



What we learn from this histogram is that the large majority of District Q's residents earn less than $35k per year, that the highest earner makes about $100k per year, and that there is a smooth transition between the lowest and highest earners.

Differential Privacy: Histograms

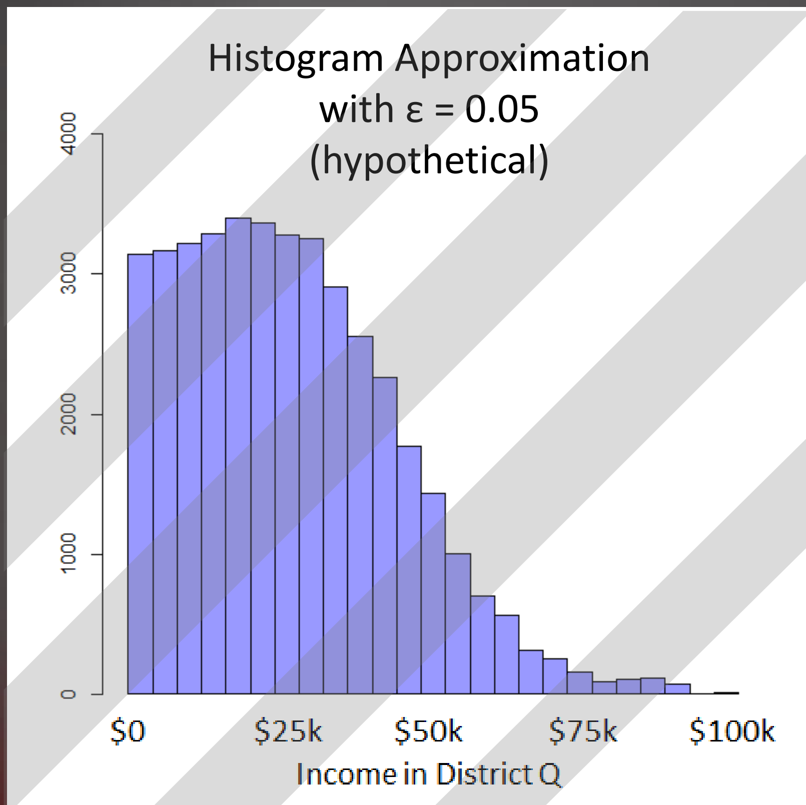
Gertrude asks the university's DP-interface for a differentially private approximation of District Q's income distribution, and sets ε = 0.01.

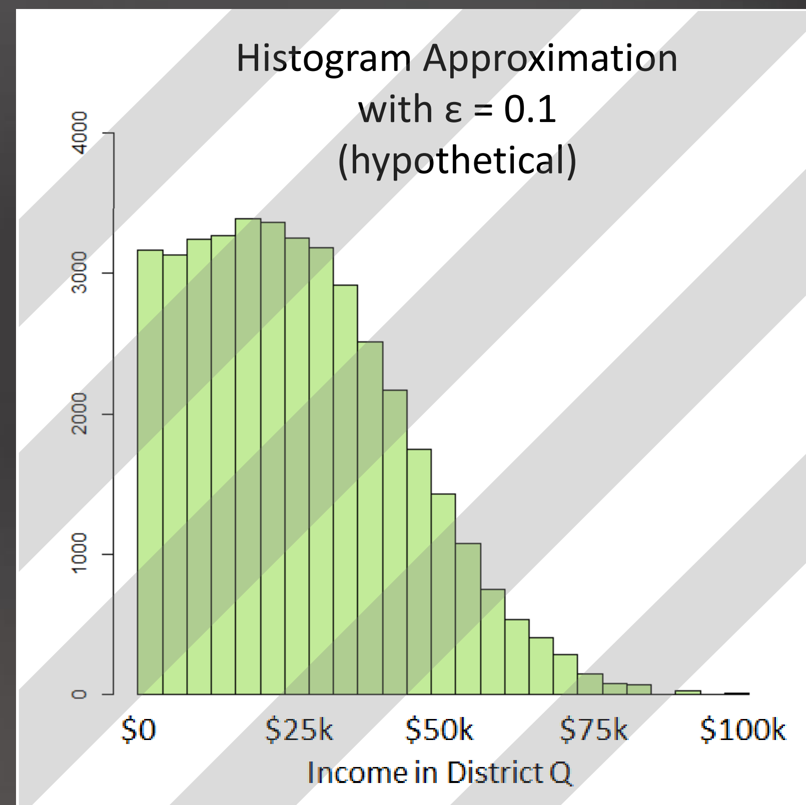


Much like the true histogram in red, this approximation suggests that District Q mainly earns less than $35k. Gertrude notices that there is apparently a spike in the number of people earning about $75k. Gertrude knows that this *may* just be a result of random noise addition.

Differential Privacy: Histograms

For educational purposes, we can look at what <u>may have happened</u> if Gertrude had used higher values for ε.

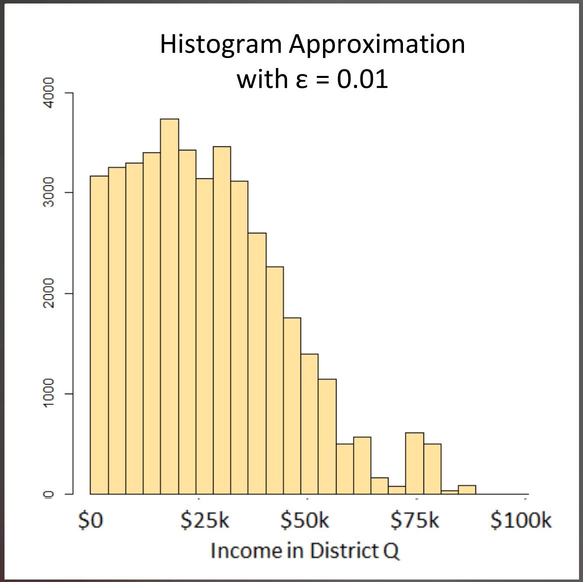


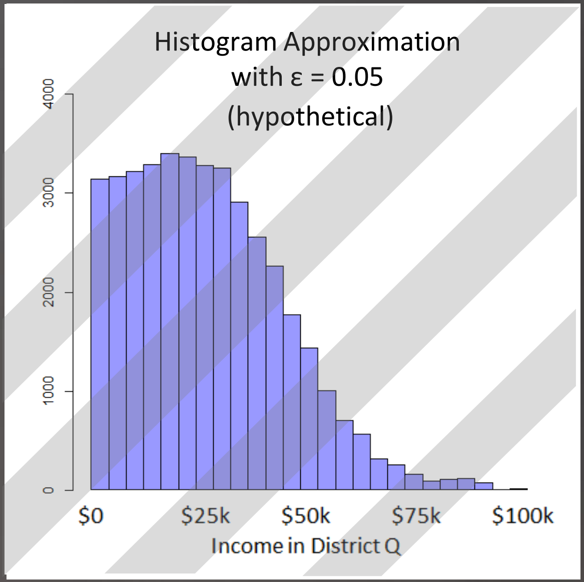


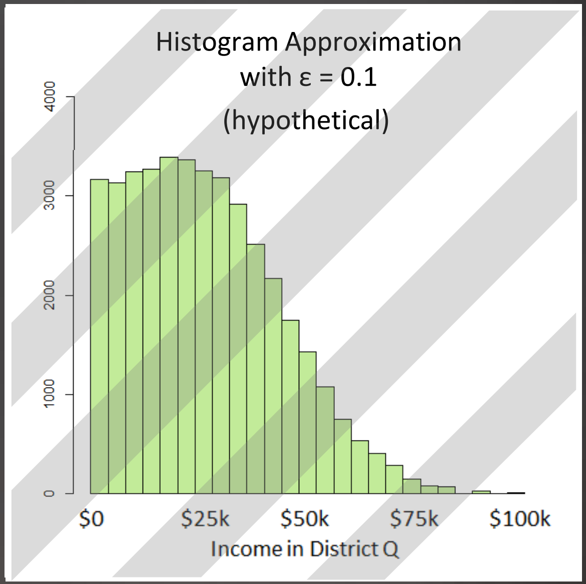
<u>Remember</u> that Gertrude, and other DP-users, would not actually be able to request more than one histogram per dataset.

Differential Privacy: Histograms

We can again view this progression in terms of the "ε-knob". Keep in mind that each of these three histograms Is the true histogram with just one instance of differentially private noise added to it. If Gertrude generated three more DP-histograms at these ε levels, she would almost certainly see different images.



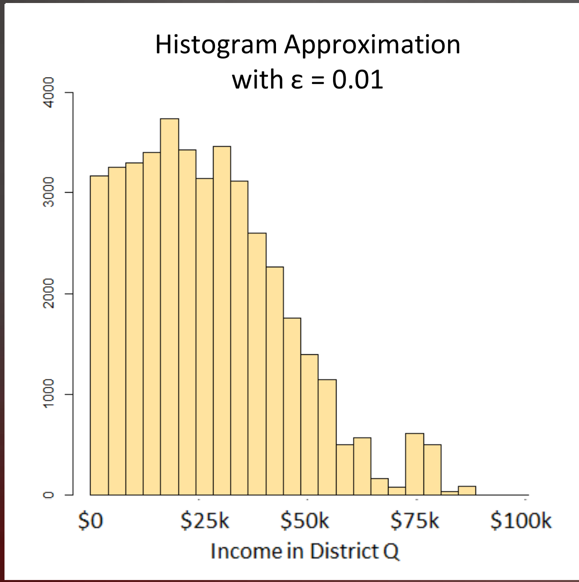


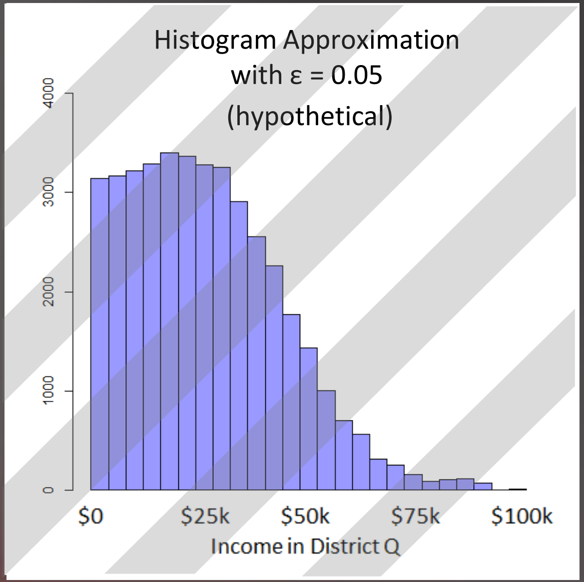


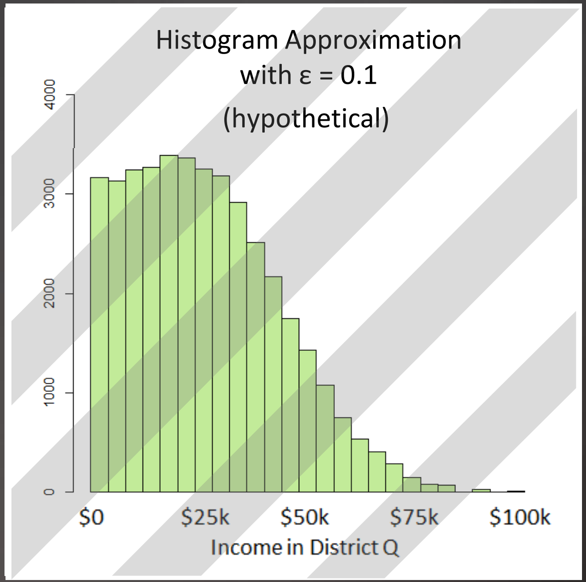
Differential Privacy: Histograms

From this, we can learn something about interpreting differentially private histograms. The key is to recognize which patterns represent the underlying data, and which patterns may just be the result of the random noise addition.

On the next slides, we'll offer an example to clarify this point.

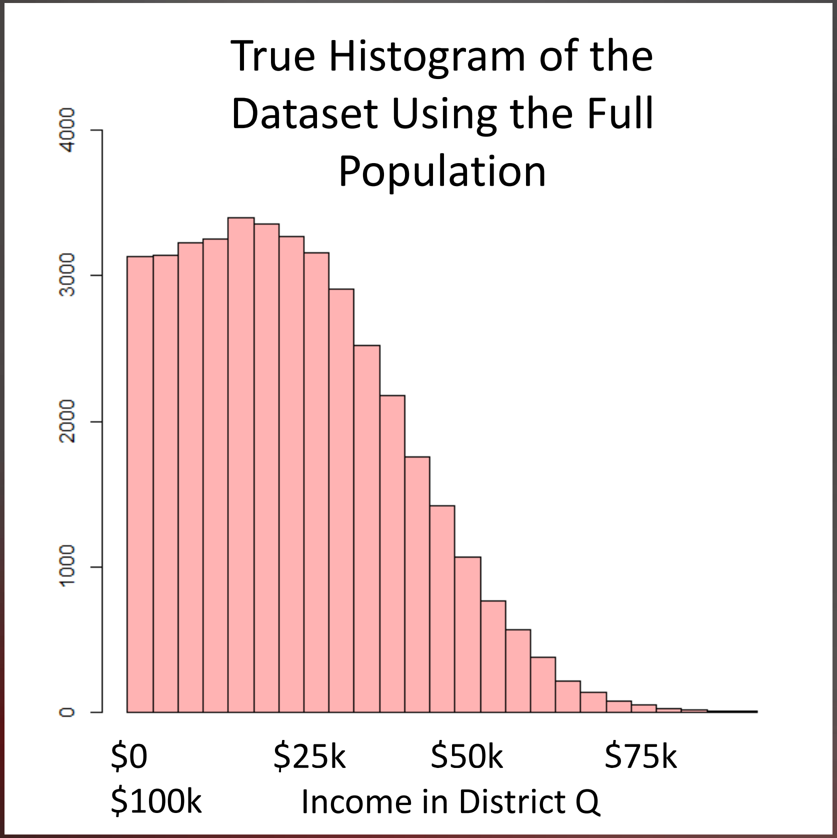


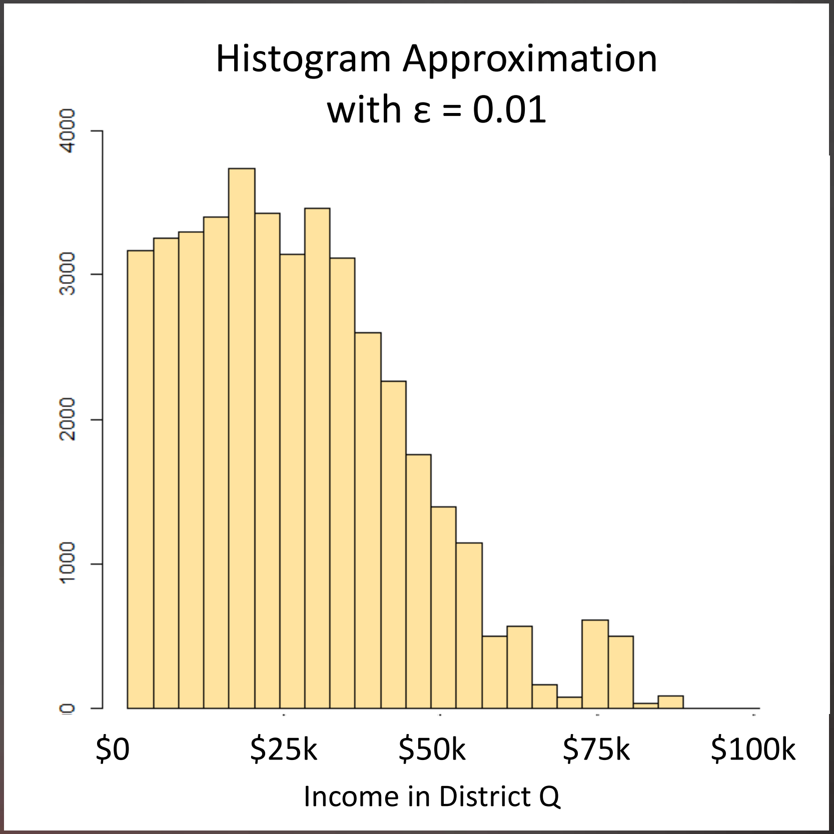




Differential Privacy: Histograms

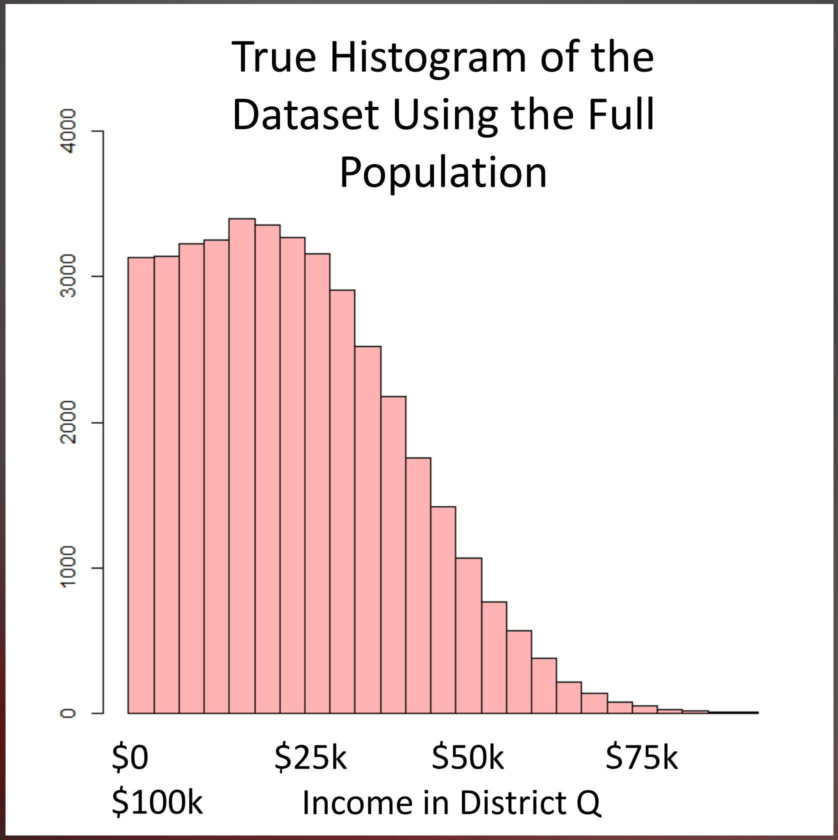
Consider the DP-histogram that Gertrude actually generated, in yellow. Keep in mind that Gertrude doesn't have access to the red (true) income distribution, and neither would any differential-privacy user.

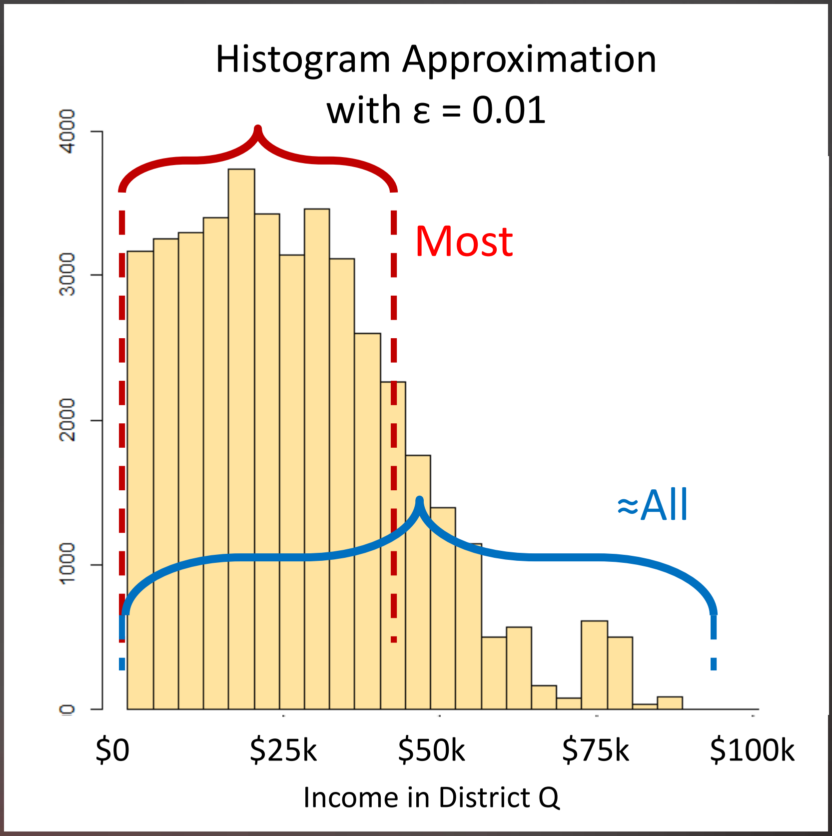




Differential Privacy: Histograms

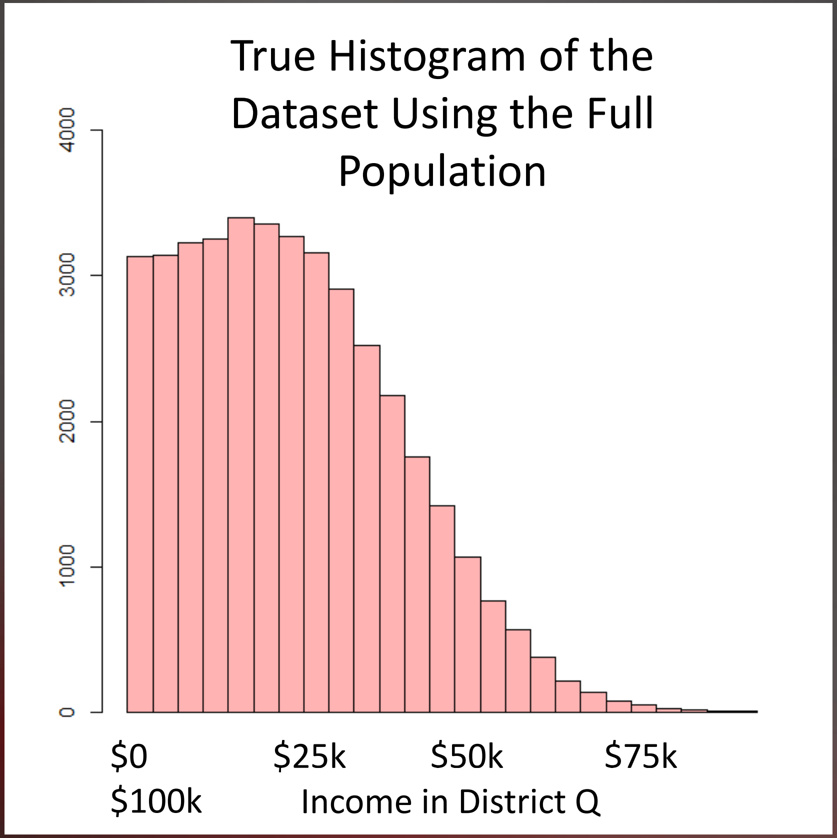
Looking at the yellow approximation, Gertrude would correctly deduce that District Q has a large proportion of residents with income between about $0K/year and $35k/year. She would also correctly assume that almost all of those surveyed earn between $0/year and $100k/year. She would also correctly assume that the income distribution has a long tail in one direction.

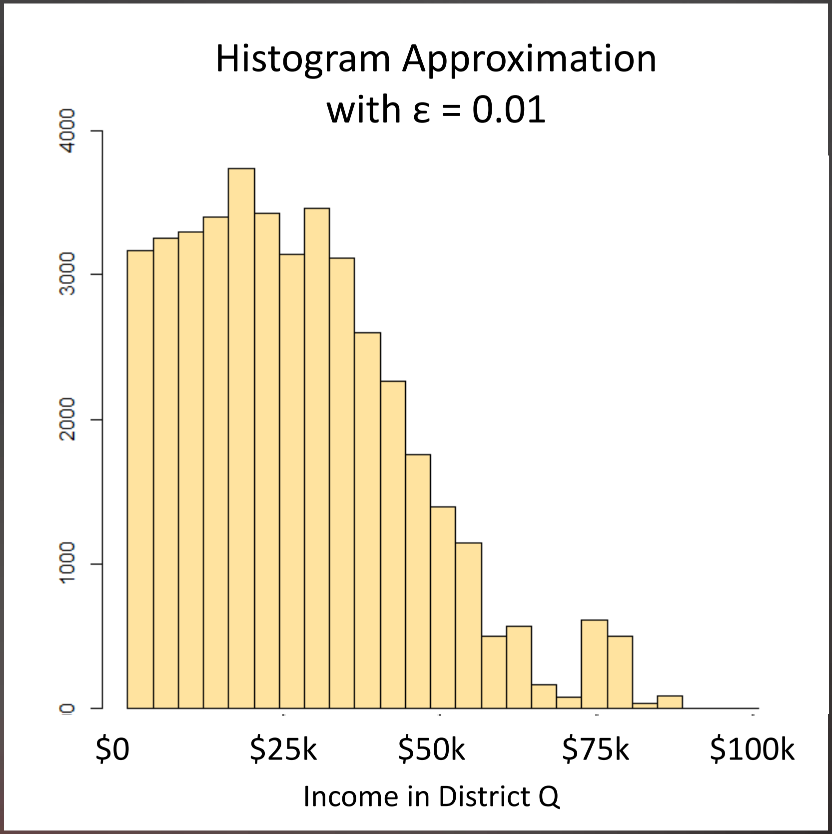




Differential Privacy: Histograms

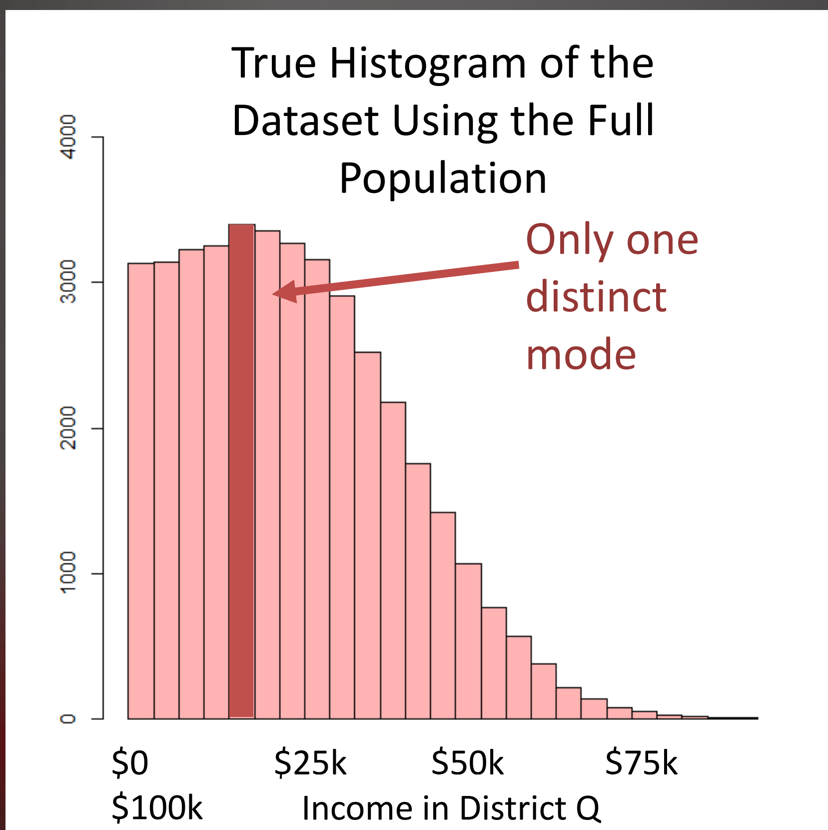
Gertrude is quite smart, and so she knows that smaller details of this histogram shouldn't be over analyzed, because they may be the effect of random noise addition. We call these effects "artificial artifacts" because they're a product of the DP-approximation process, not the underlying data. We'll now look at a few of these artificial artifacts.

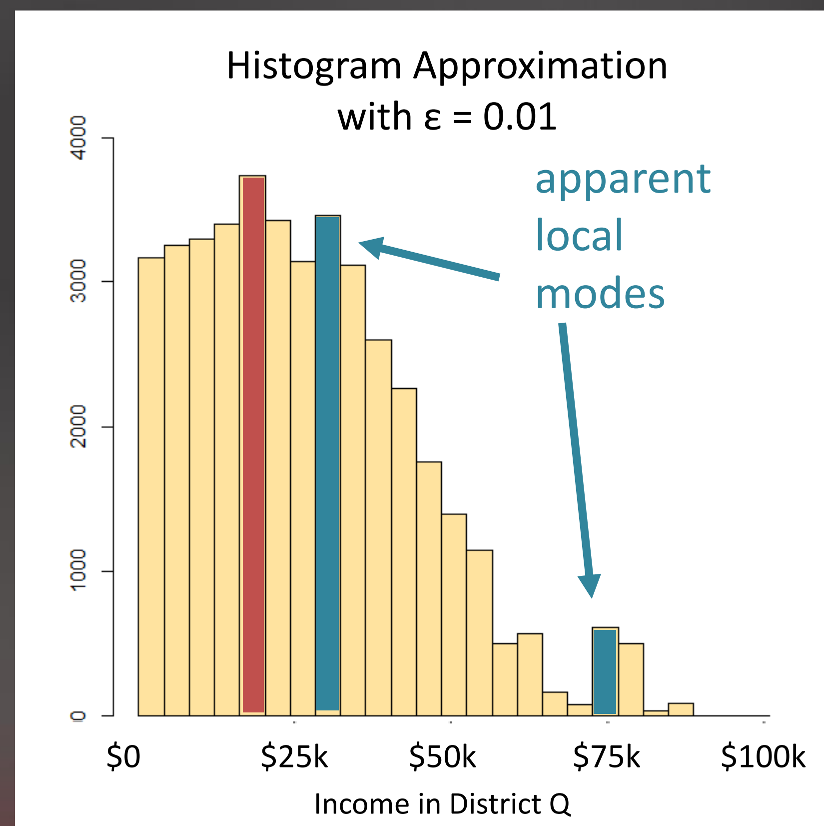




Differential Privacy: Histograms

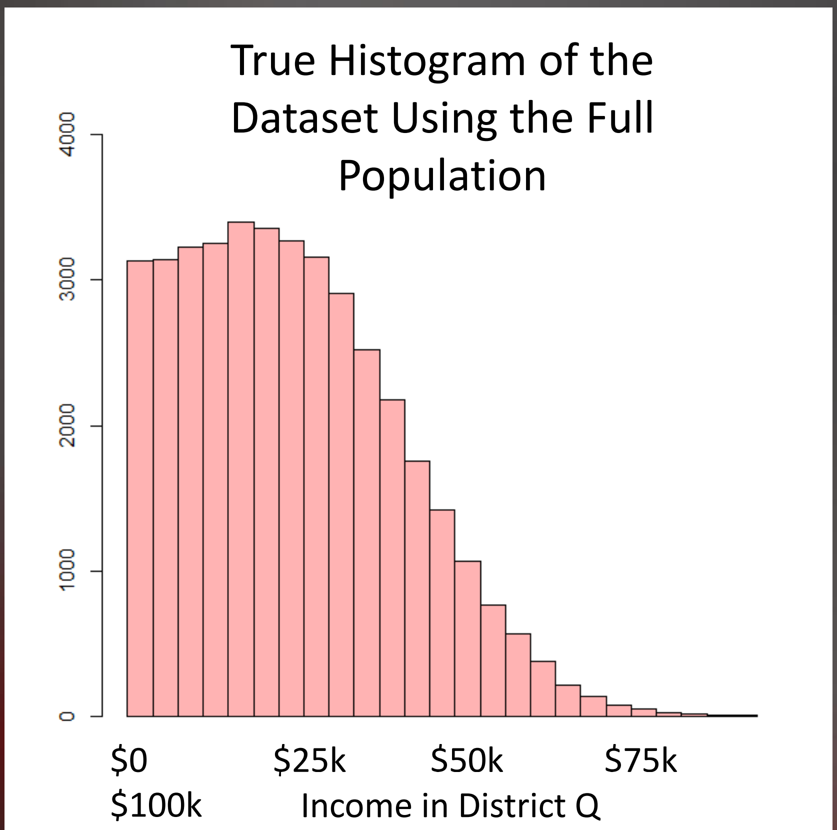
First, Gertrude knows that she does not know the exact mode of this data with certainly. Histogram bars similar in height can surpass each other when random noise is introduced, seemingly switching what the mode is. We see a similar common situation here, with apparent local modes introduced. However, Gertrude can feel confident that the actual mode lies in one of the few tallest bins.

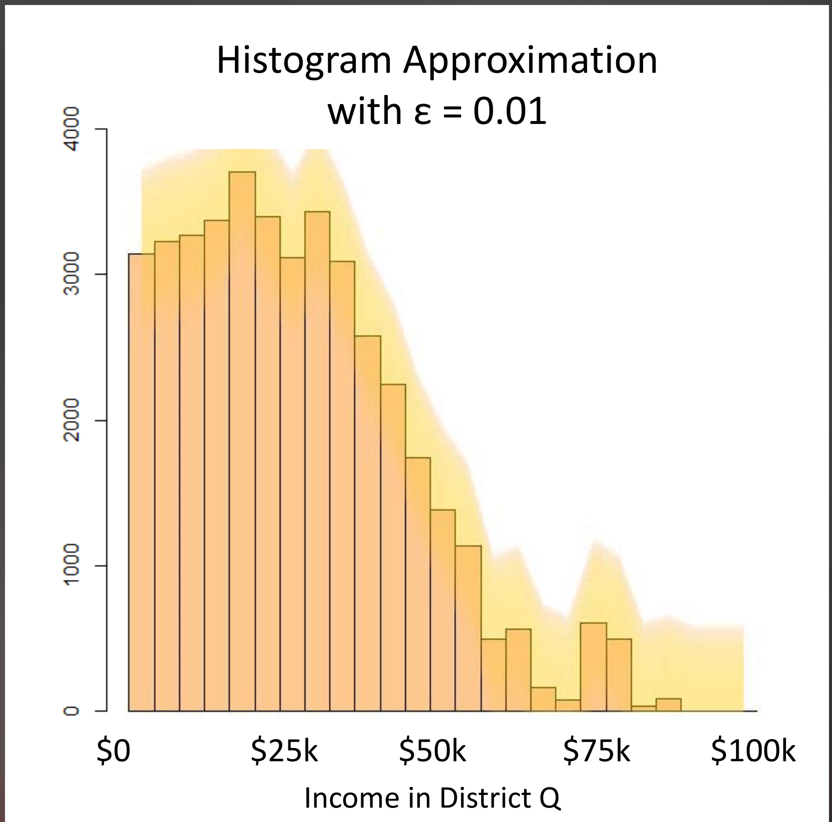




Differential Privacy: Histograms

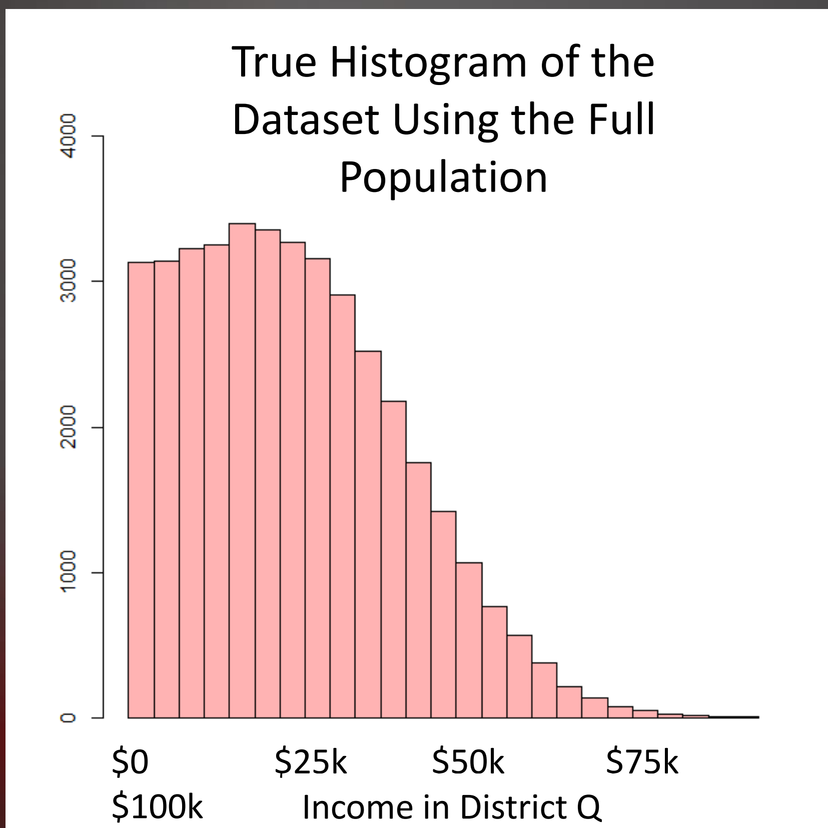
Knowing that random noise can cause similar bins to become taller or shorter than their neighbors, Gertrude knows to interpret the DP-histogram as though the tops of the bins may fall within an enveloping sleeve, as pictured below. Small skips and jumps within the sleeve may be the effect of DP-noise, artificial artifacts, but the sleeve itself shows us the trends we want to see.

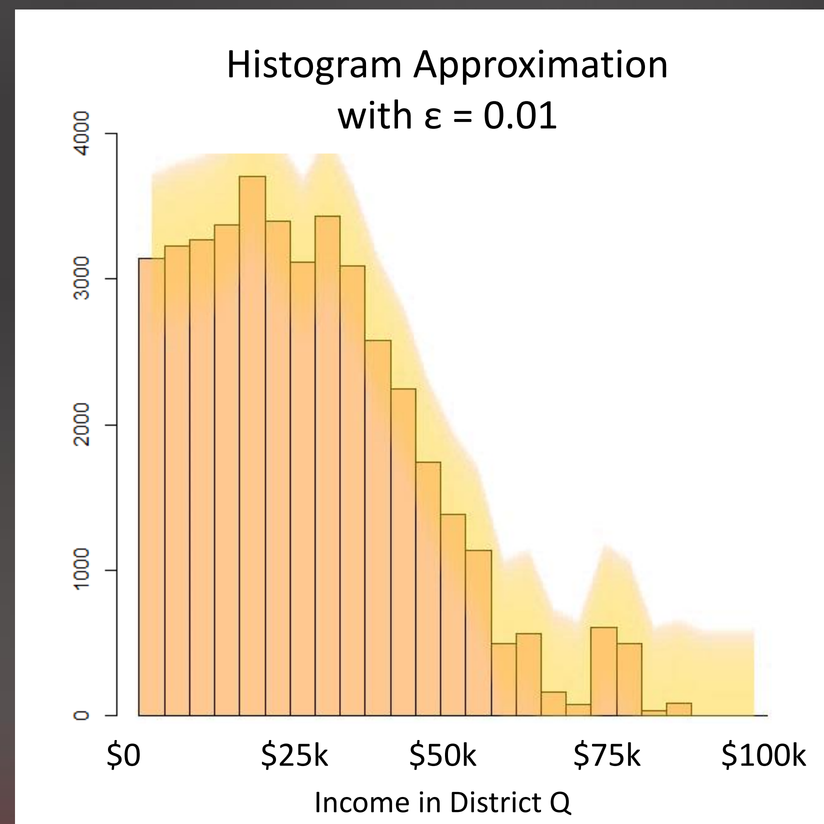




Differential Privacy: Histograms

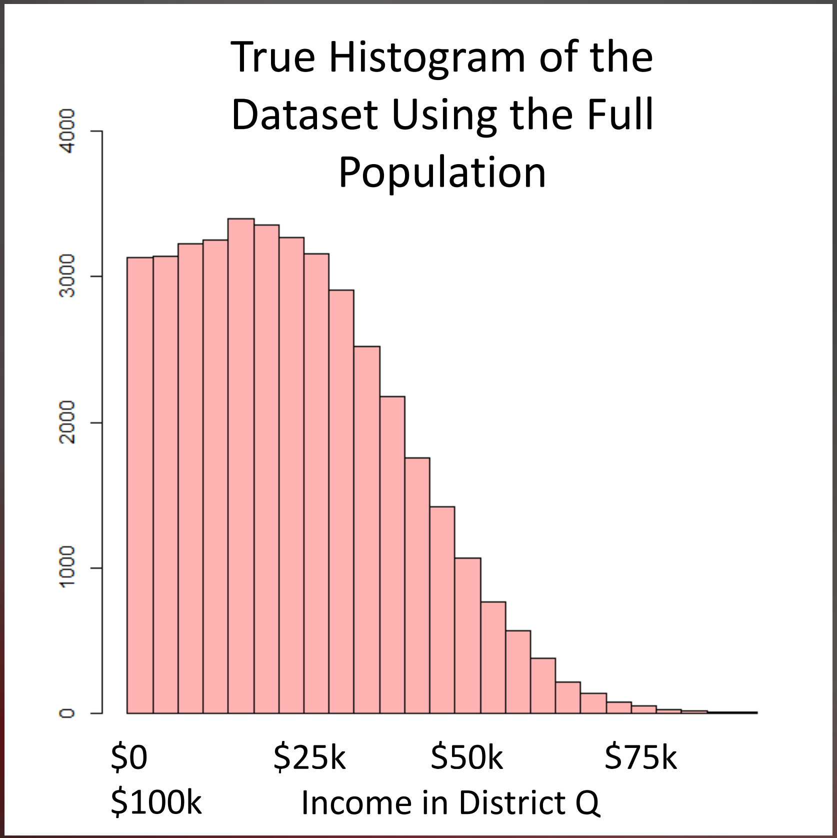
In a way, this sleeve represents the privacy that DP provides. It is mathematically calculated such that the midpoint of the top of each bin of the *actual* (non-DP) histogram falls within the yellow sleeve. Said another way, the blurry sleeve represents a 95% confidence interval on each bin. Certain DP interfaces provide this blurry confidence-interval sleeve to users.

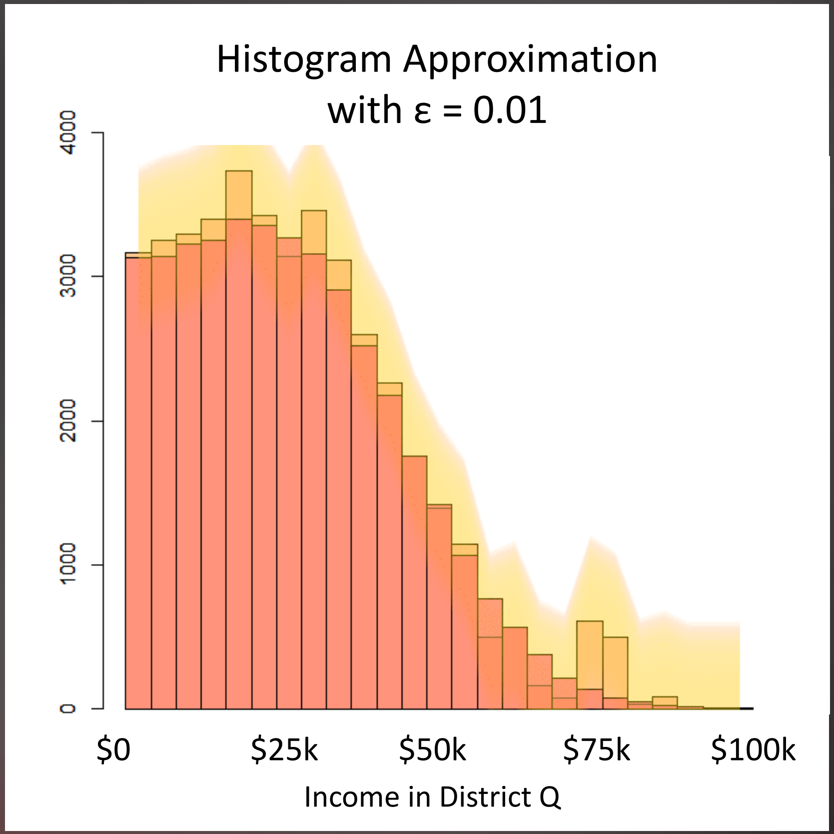




Differential Privacy: Histograms

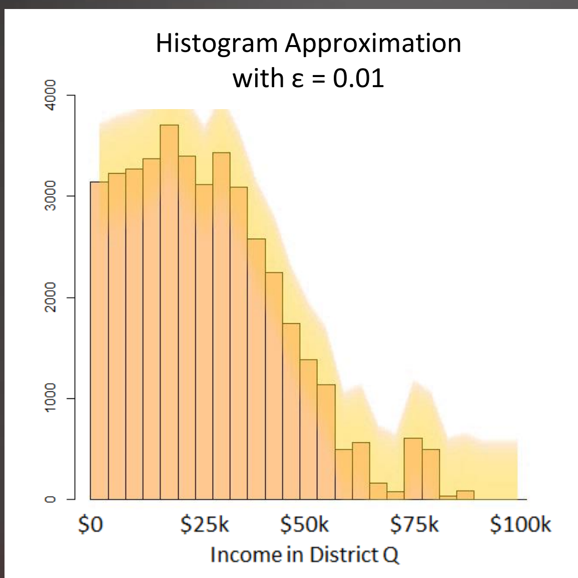
By using the outline of the true histogram, we can see that it fits neatly within the sleeve, though it clearly does not match the yellow DP-Histogram beneath. In the next slide, we'll refine this sleeve for more detail.

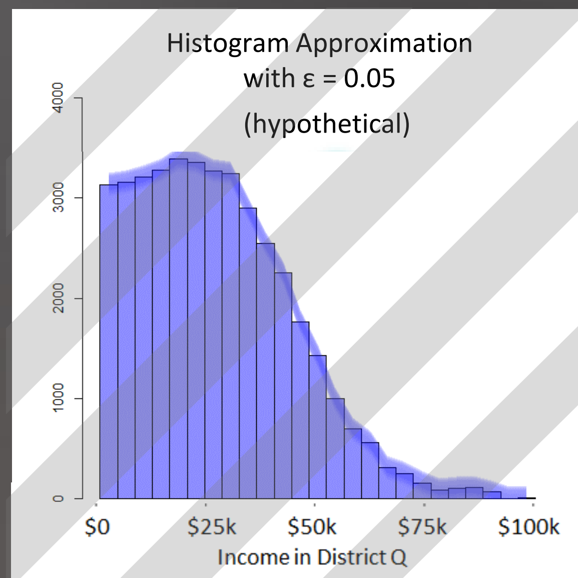


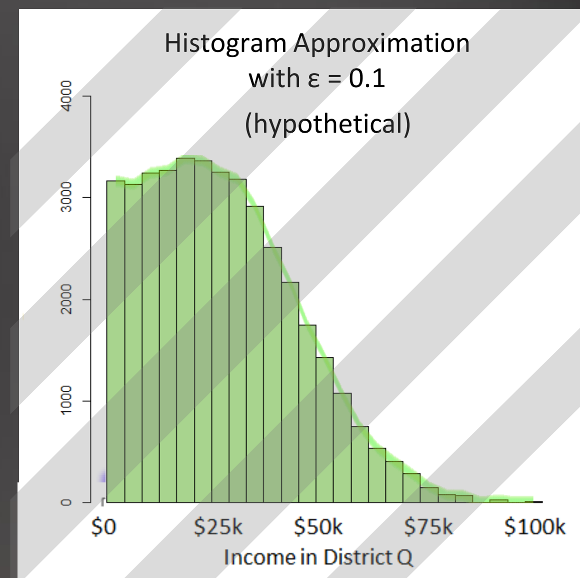


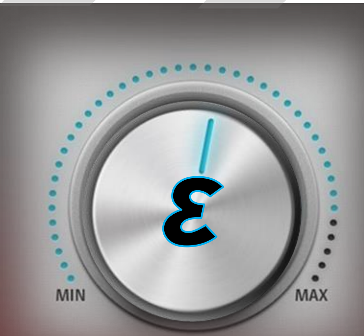
Differential Privacy: Histograms

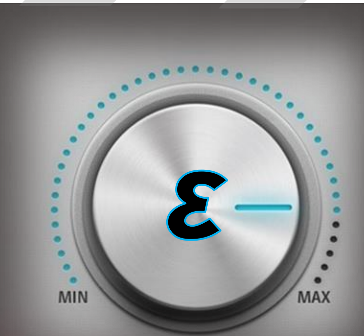
With higher ε values, we can imagine this sleeve getting thinner, revealing more detail. Notice that the confidence intervals on the green histogram are especially thin, reflecting the higher accuracy gained with higher ε values.





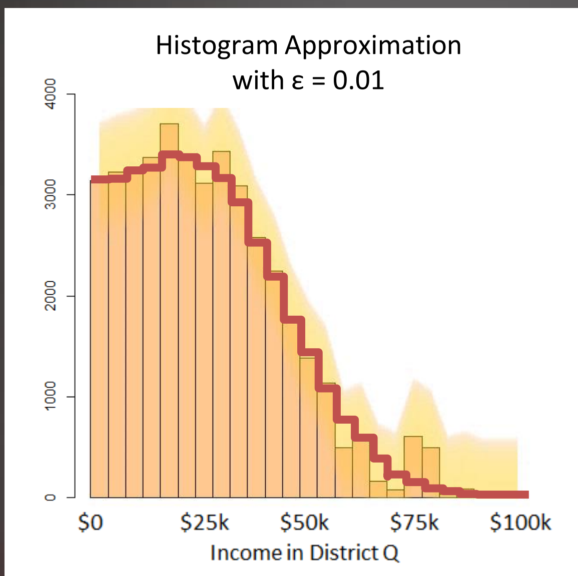


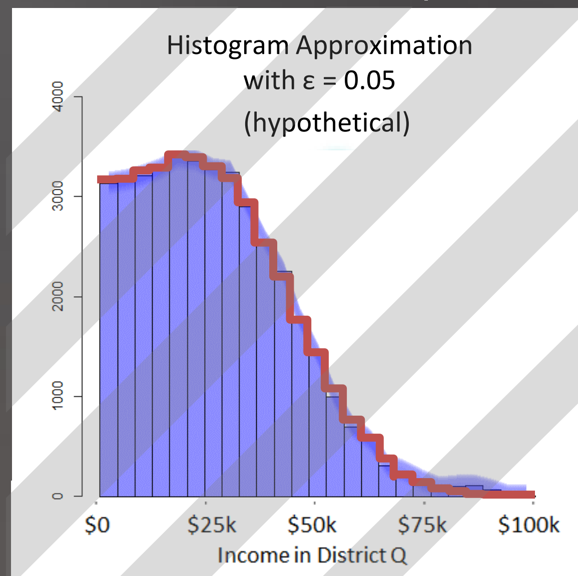


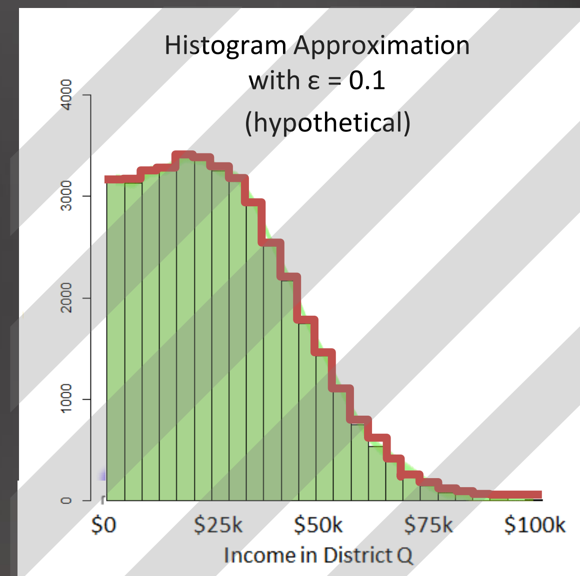


Differential Privacy: Histograms

Lastly, by adding in the outline of the true histogram, we can see how the shrinking sleeves contain the true underlying distribution, and how approximations made with higher ε generally resemble the true histogram more closely.

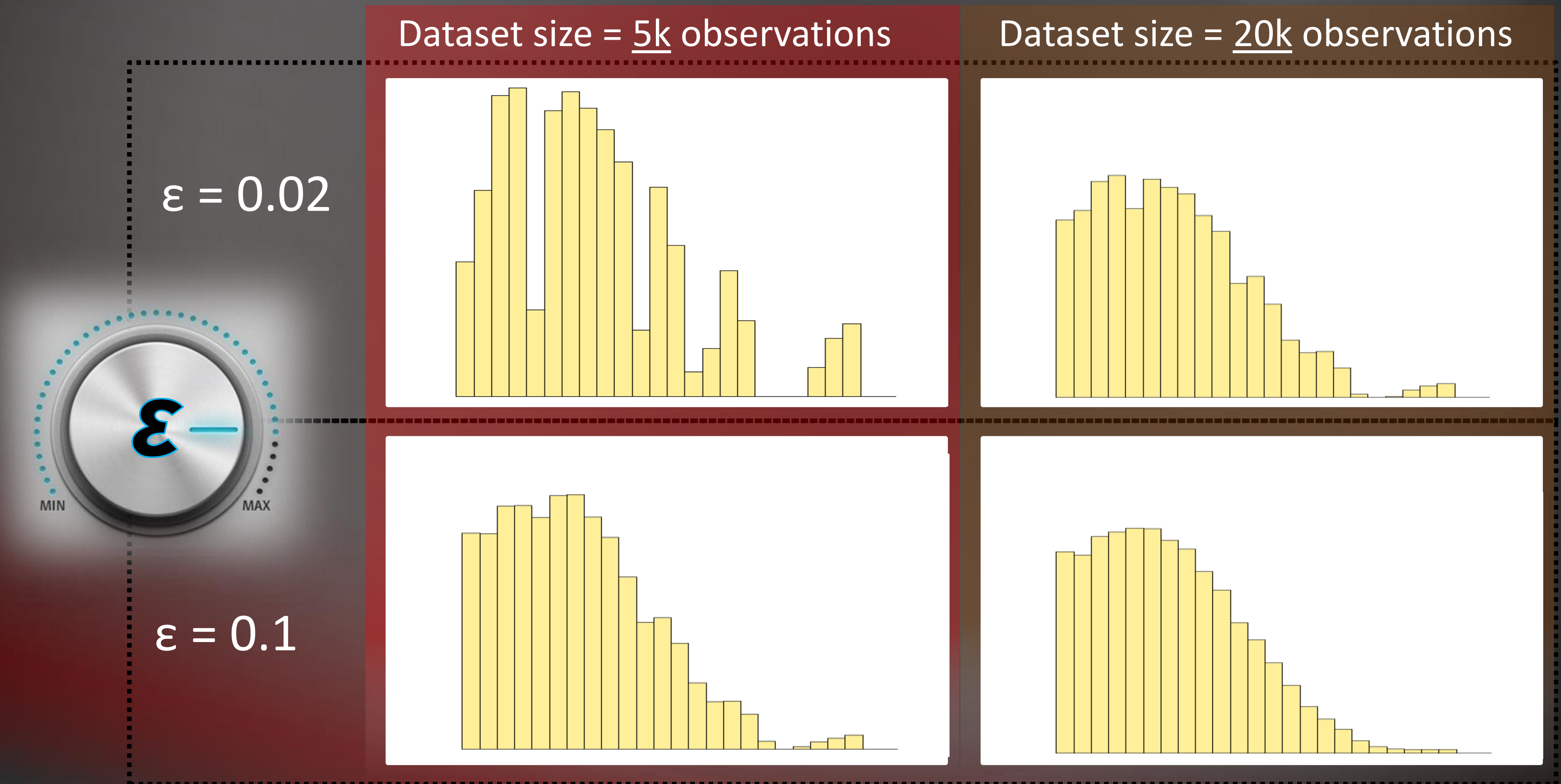






ε & n in DP-histograms

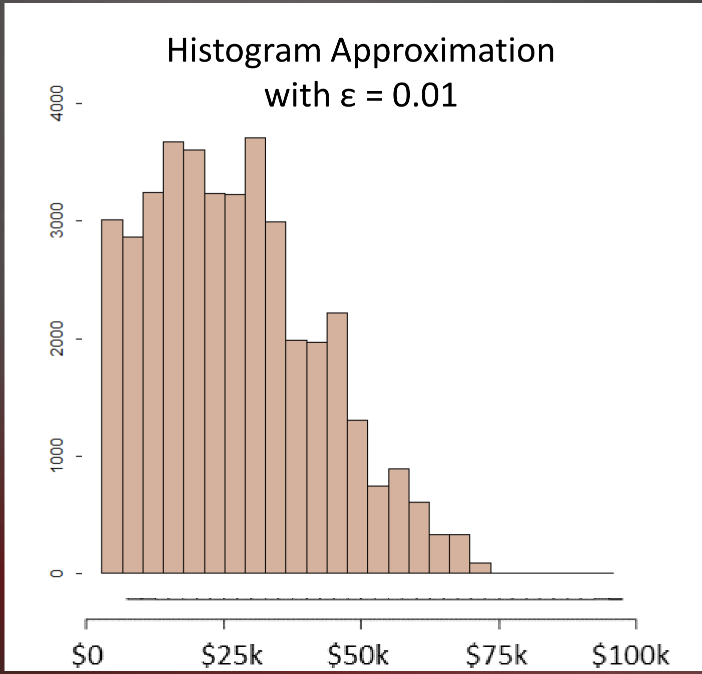
ε is not the only important parameter when constructing DP-histograms. In DP-histograms, we can also consider dataset size. The larger the set, the lower the effect of random noise. Below we have two examples of this effect.

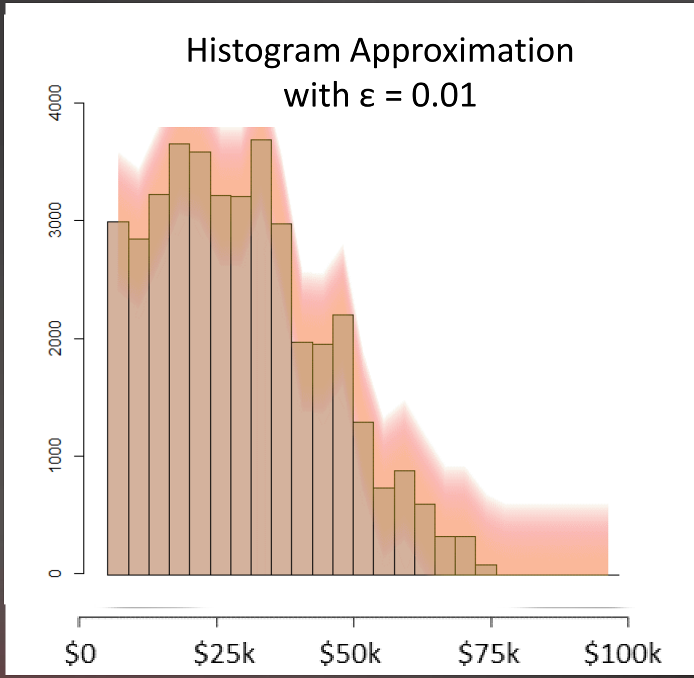


Differential Privacy: Histograms

Finally, after viewing a DP-histogram Gertrude decides she wants to access the university's District Q income survey data for her research.

The university also surveyed District D. Gertrude wants to determine if she should seek full access to that data as well. She runs a DP-approximation with ε = 0.25 and gets the result below. District D seems similar, so she decides to seek access.





Section 4.3

DIFFERENTIAL PRIVACY: APPLICATION

Means | Histograms | CDFs

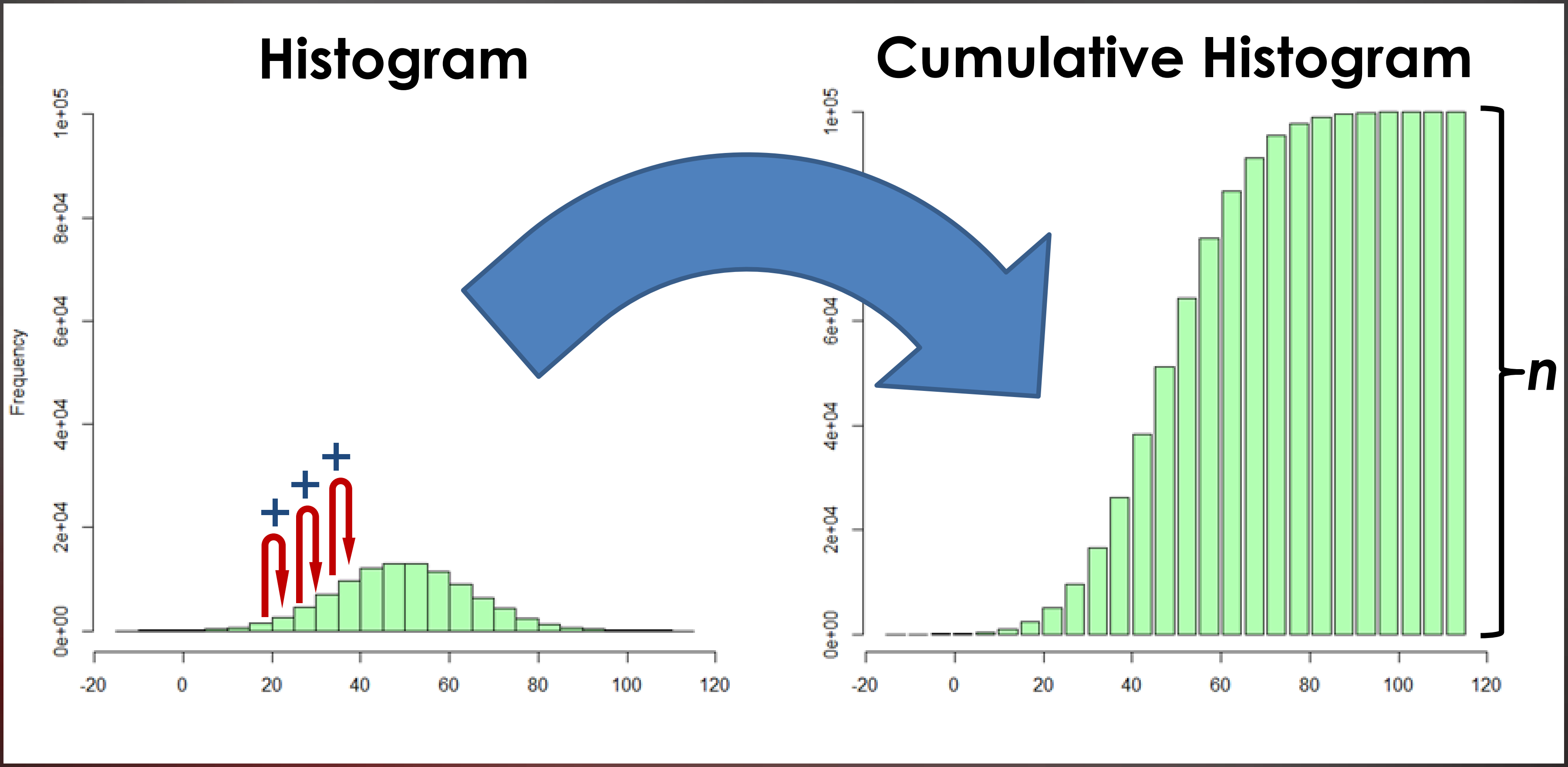
Differential Privacy: CDF

With this proper understanding of differentially private histograms, we can now move on to reading differentially private Cumulative Distribution Functions, or CDFs.

Interpreting differentially private CDFs is very similar to interpreting differentially private histograms.

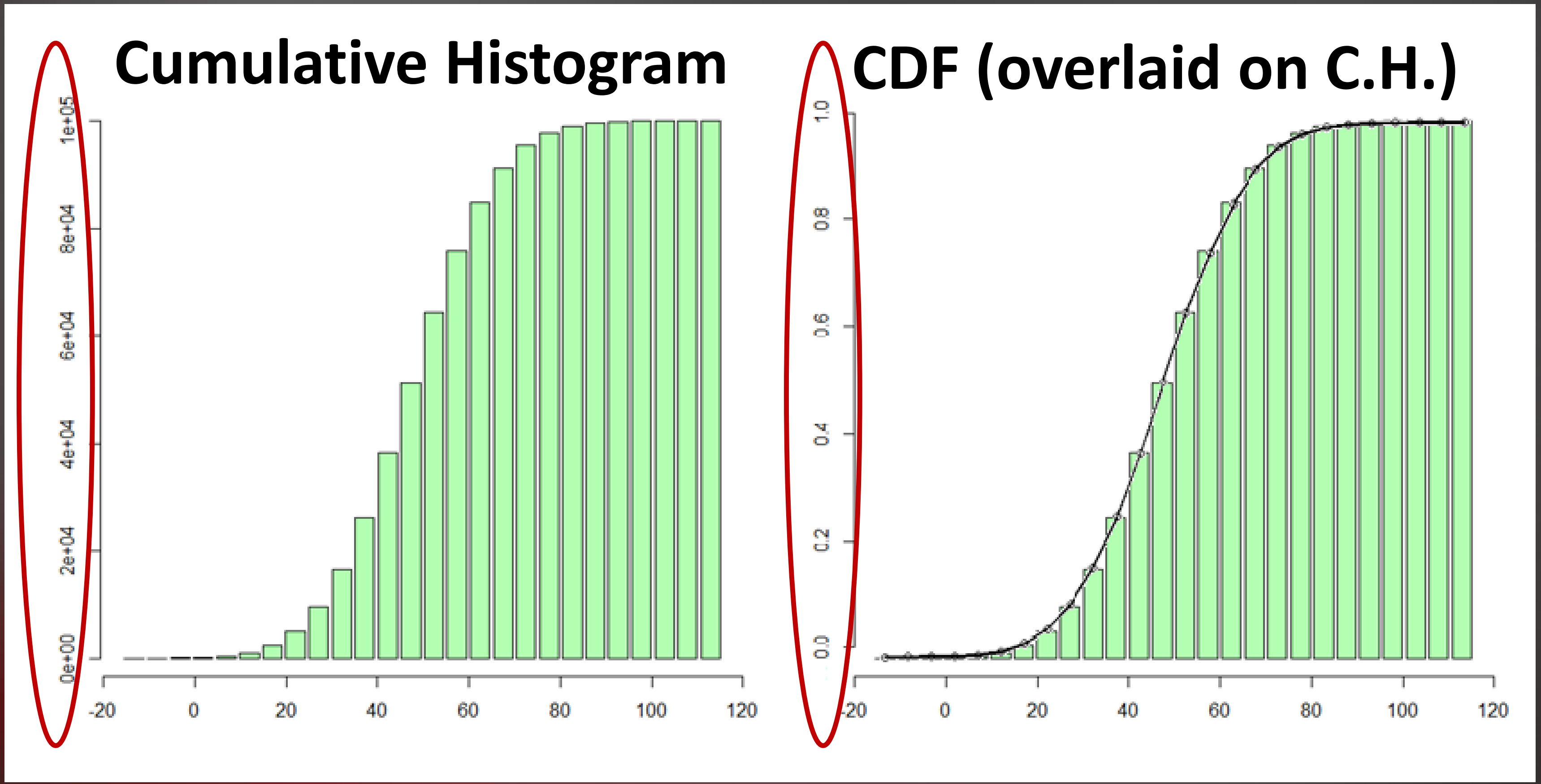
Recap: Cumulative Density Functions

Recall that CDFs are made by first adding the bars of a histogram gets us a cumulative histogram...



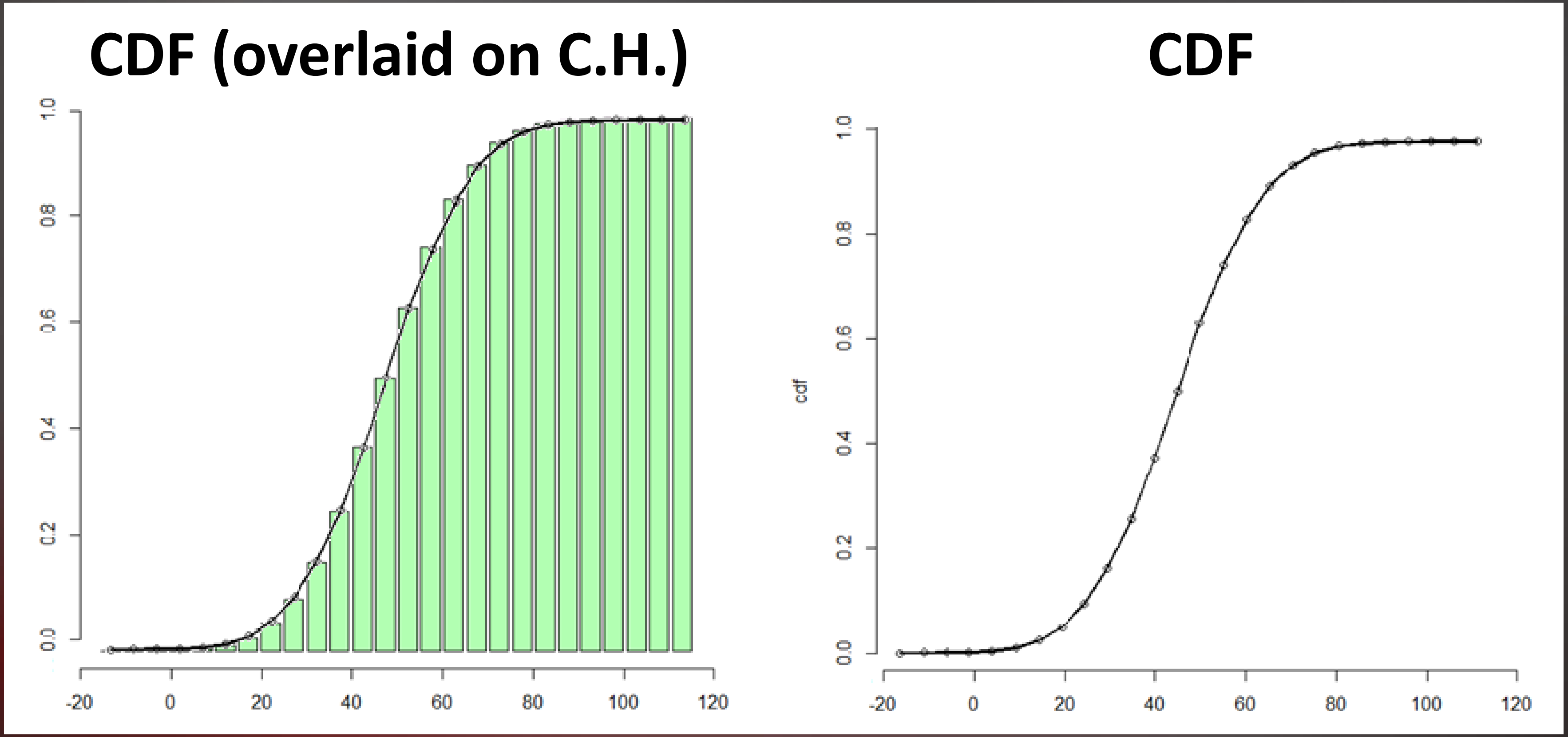
Recap: Cumulative Density Functions

…and then by dividing by the number of observations in the data….



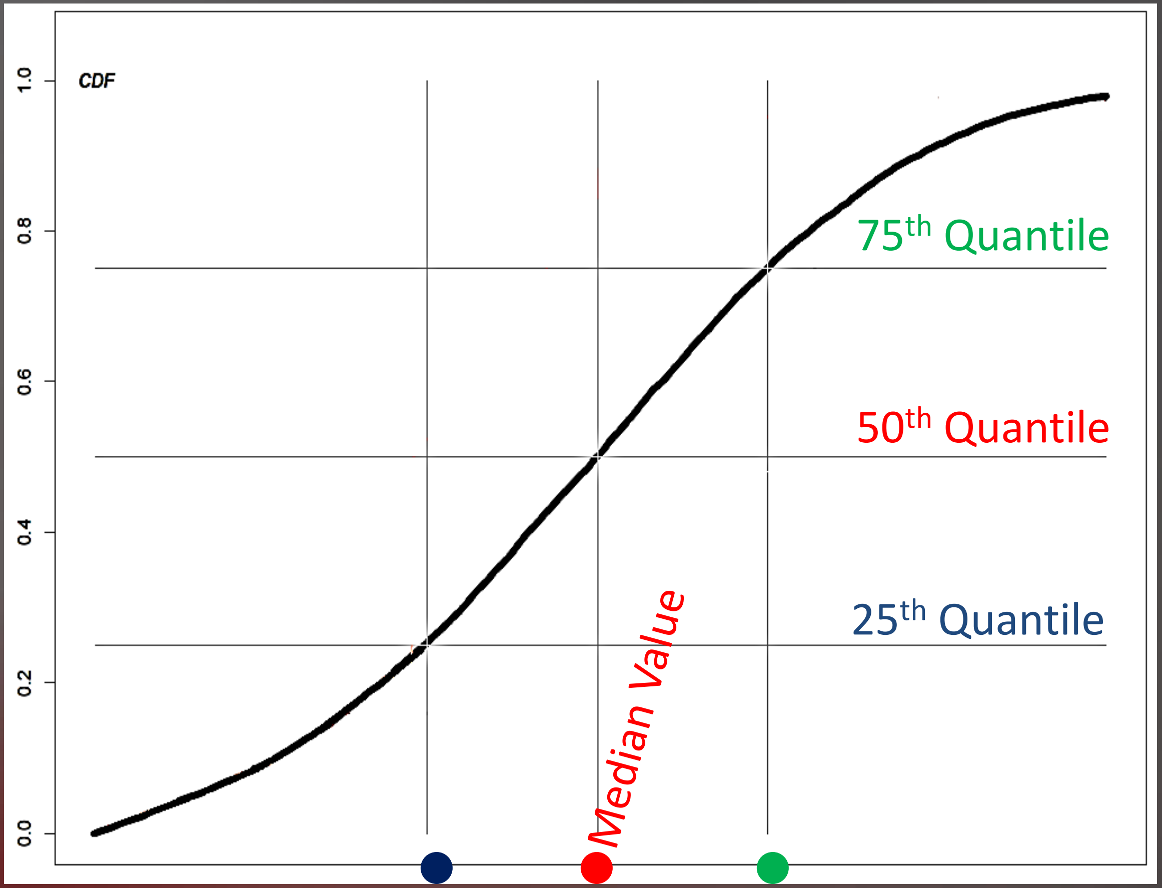
Recap: Cumulative Density Functions

… and then by removing the underlying bars used to compute CDFs.



Recap: Cumulative Density Functions

Also recall that CDFs can show us distributions and quantiles, as shown below.

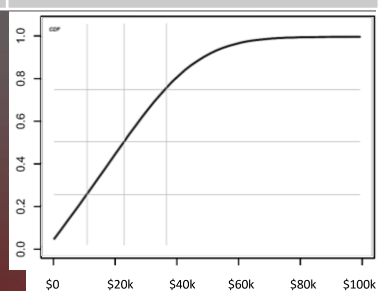


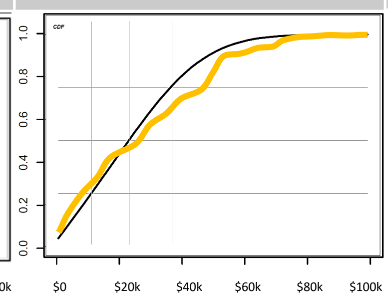
Recap: Sampling error in a CDF

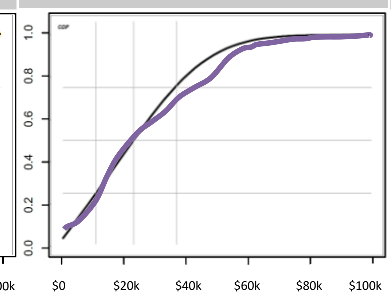
Lastly, here are the earlier results from examining sampling error in CDFs. As we saw, the sampled-approximation of the true CDFs caused the curve to separate form the true CDF, and caused the apparent median to shift. As we'll see, differential privacy can have a very similar effect.

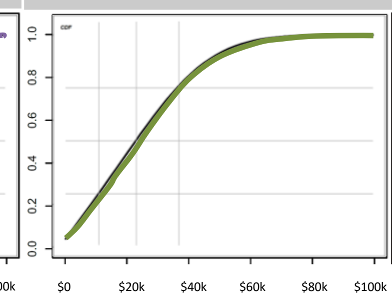
Researcher		**Gertrude**	**Colleague 1**	**Colleague 2**
Sample Size	**(40,000)**	**500**	**2000**	**10,000**
Median in USD	23,500	26,034	22,309	23,870
Difference*	(0)	+2534	-1191	+370

*the empirical median found by each researcher minus the true median, 23,500.



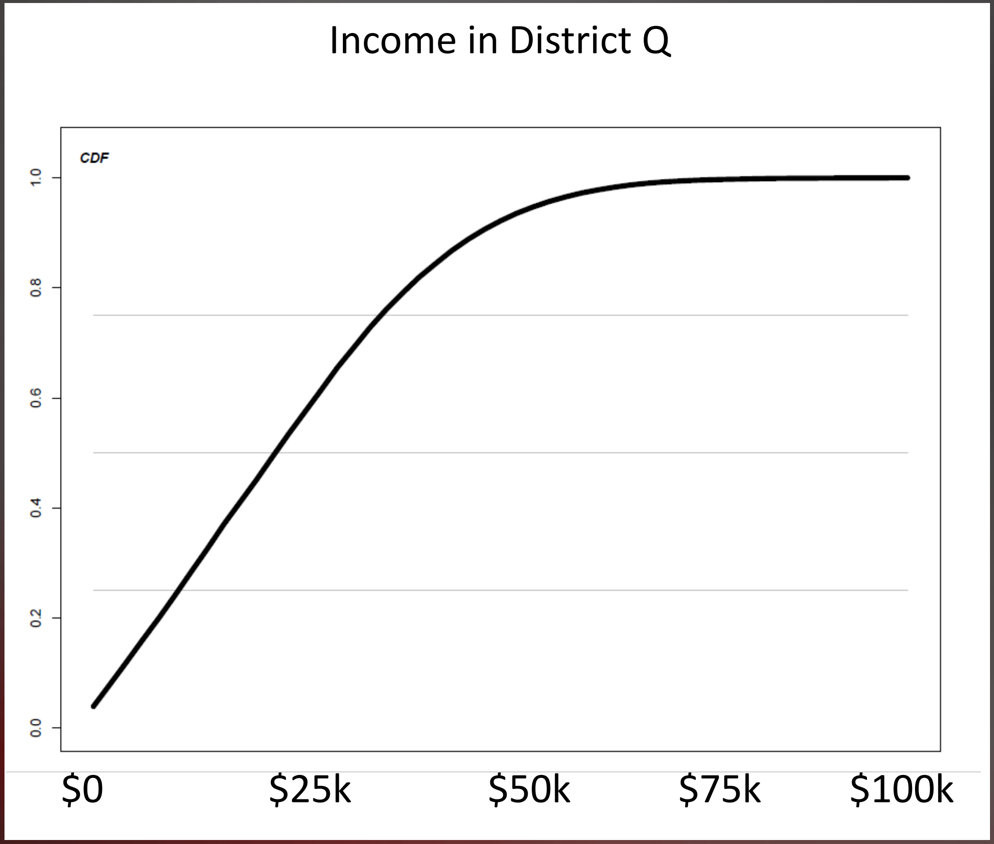






Differential Privacy in CDFs

First, we'll look at this CDF. This CDF is made from the university's data on income distribution in District Q. Researchers using the differential privacy interface cannot see this CDF.

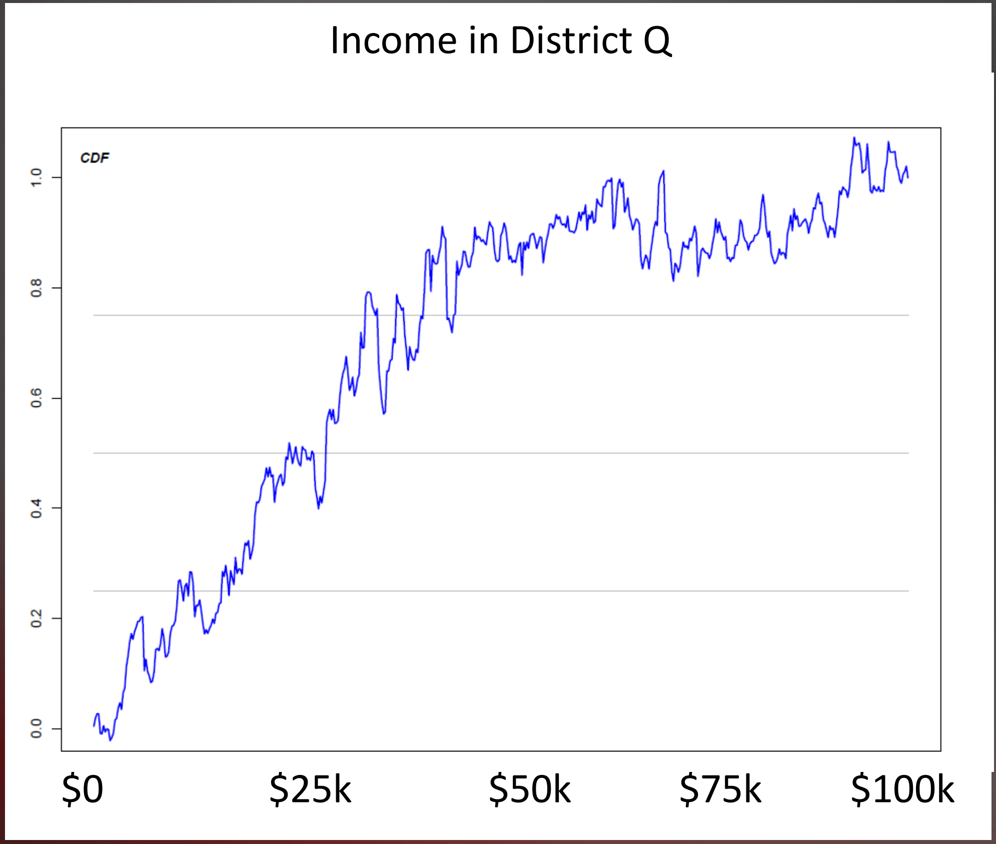


As we saw in the histogram, the income is smoothly distributed. Here, we see that the median is $50k, and the 25th and 75th percentiles are $35k and $65k respectively.

We'll continue with Gertrude, and observe how she analyzes a few differentially private approximations of this CDF.

Differential Privacy in CDFs

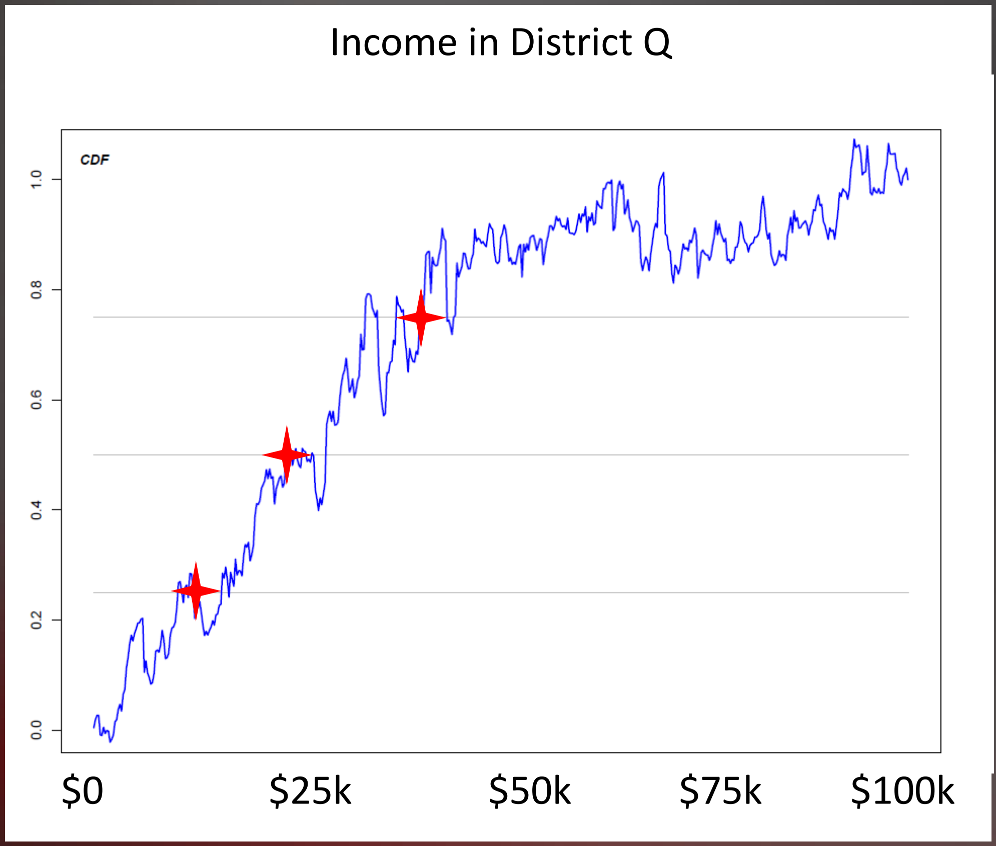
In the graph below, we see Gertrude's differentially private approximation of a CDF. This blue CDF was made with ε = 0.01. Gertrude first notices that the income distribution is fairly equal overall. She also notices that this CDF is very jagged.



The clearest effect of differential privacy's random noise is that many parts of this CDF show negative probability by dipping downward. This is a clue that the jaggedness we see is from random noise, and not the underlying distribution.

Differential Privacy in CDFs

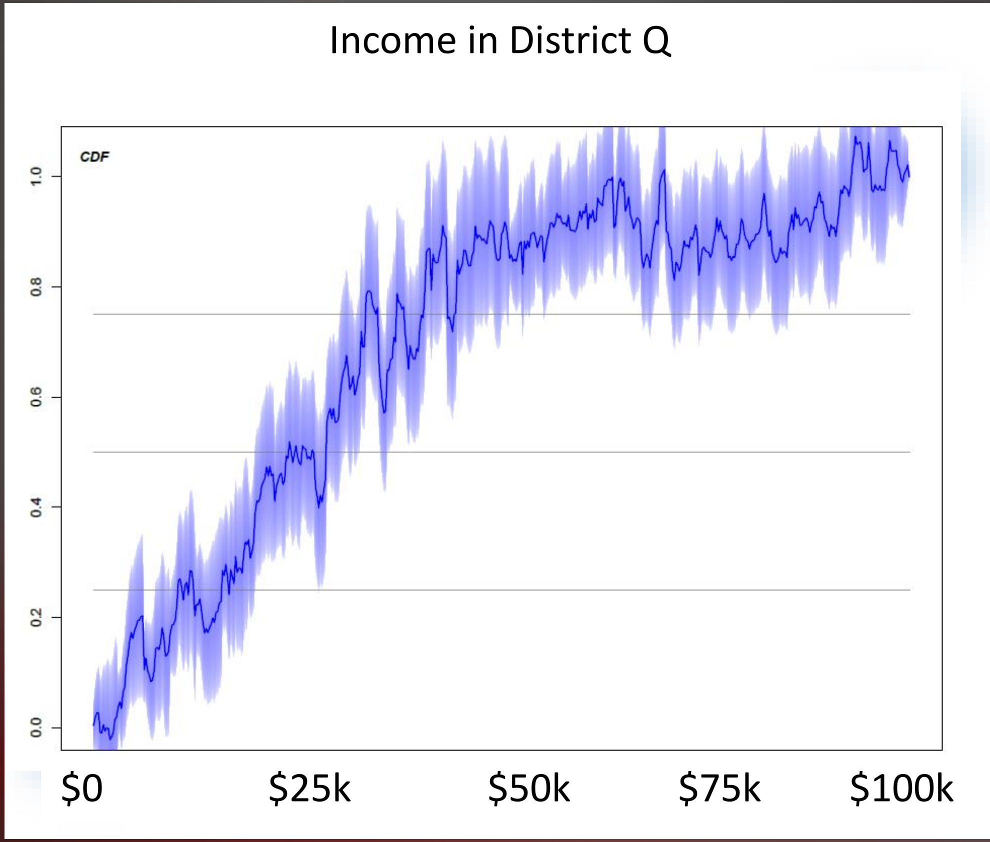
CDFs are commonly used to locate medians and quantiles. We can see here that differential privacy has shifted our median and key quantiles slightly.



Percentile	CDF from Actual Data	DP-CDF
25th	$11k	$14k
50th	$21k	$23k
75th	$32.5k	$37.5K

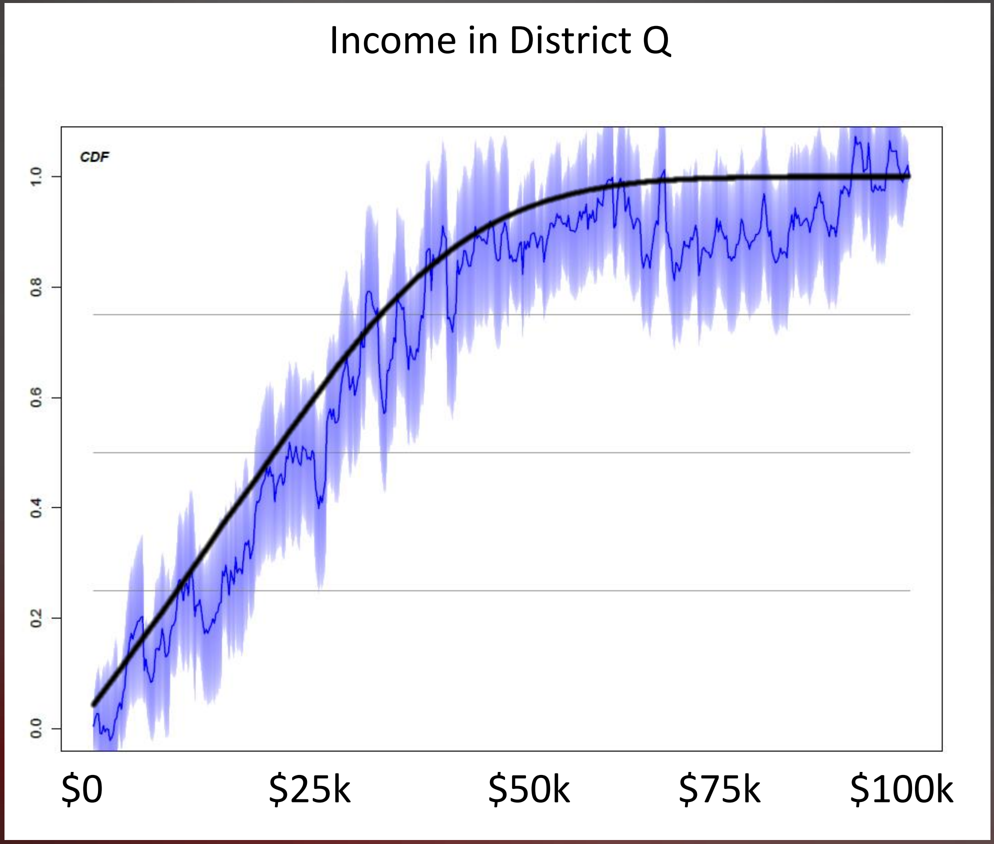
Differential Privacy in CDFs

As with the histograms, it's helpful to envision a sleeve or cloud around the CDF. This sleeve depicts the 95% confidence interval of each point. We can see that most of the apparent jaggedness of the CDF is contained within the sleeve.



Differential Privacy in CDFs

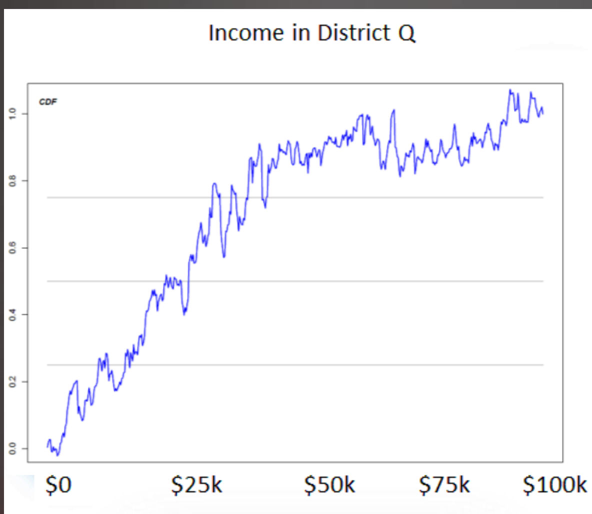
When we overlay the true CDF, we can see that roughly 95% of it fits within this sleeve, as expected. Using this true CDF for reference, we can also see more clearly how differentially private noise affected the result.



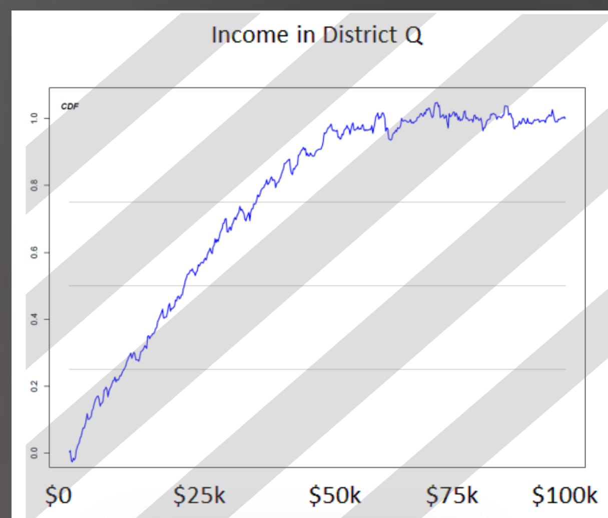
Differential Privacy in CDFs

Now let's return to the ε knob. Below we see the DP-CDF that Gertrude was given, as well as two hypothetical DP-CDFs with different values for ε.

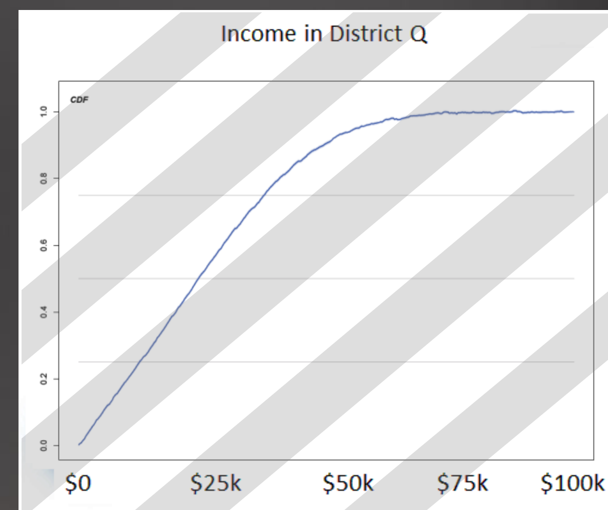
ε = 0.01



ε = 0.025



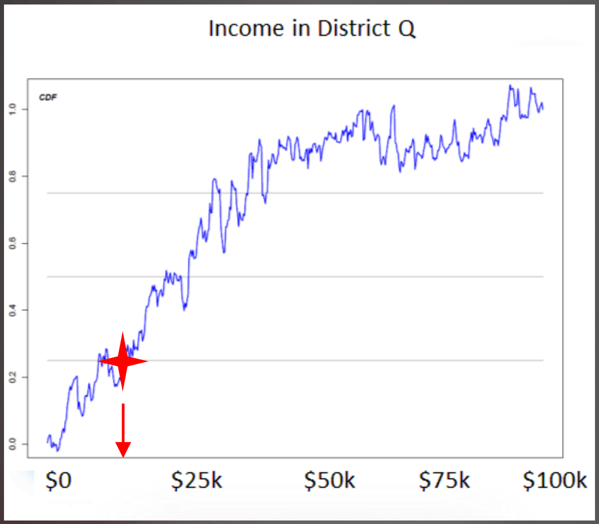
ε = 0.25



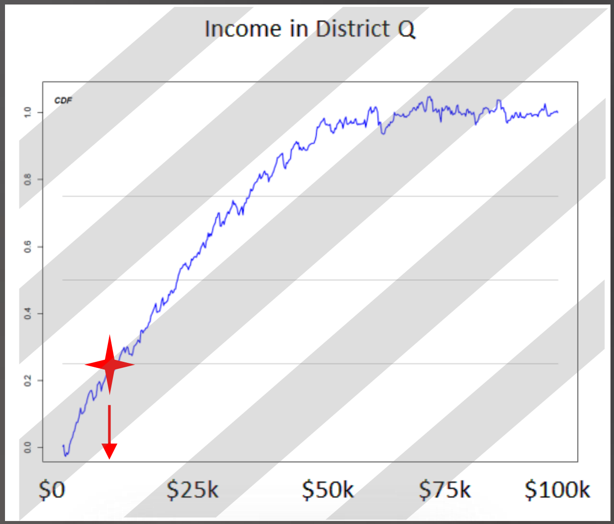
Differential Privacy in CDFs

Per usual, Gertrude knows better than to interpret DP-CDFs as exact CDFs. For example, notice that if we **incorrectly** treat these DP-CDFs as if they are exact, our apparent 25th percentile may be read as exactly $14K, $10k, or $11k.

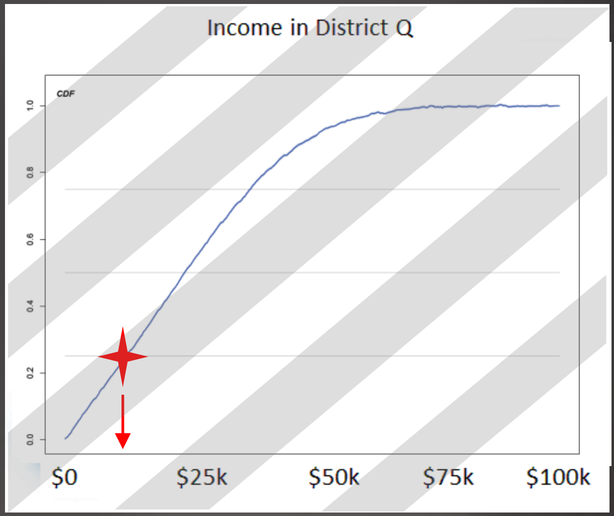
ε = 0.01



ε = 0.025



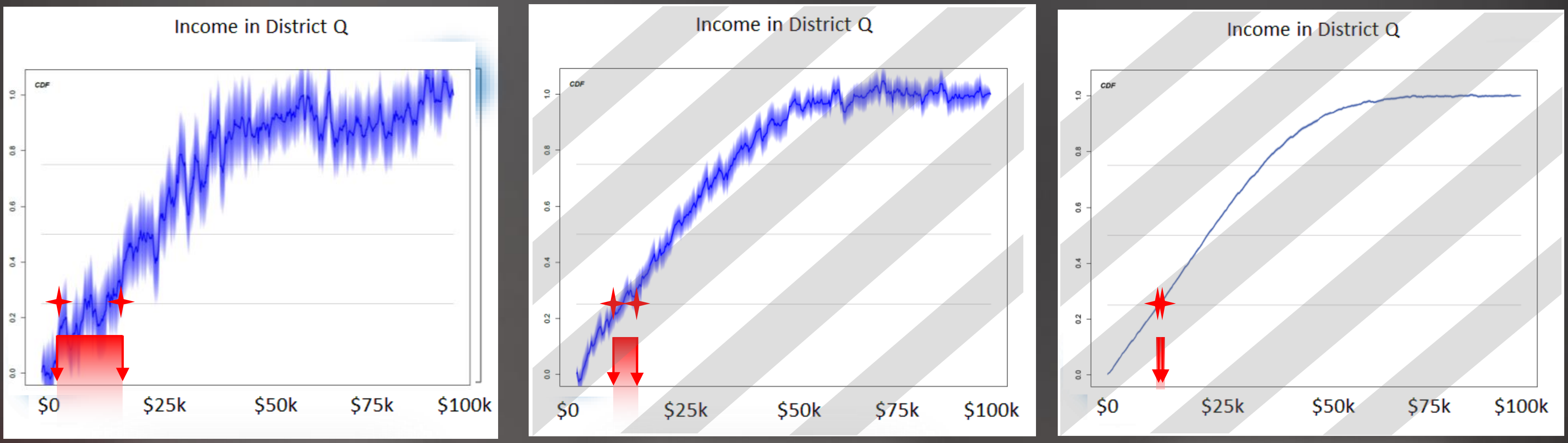
ε = 0.25



Differential Privacy in CDFs

Instead, Gertrude smartly thinks about sleeves around these DP-CDFs. This way, she **correctly** interprets that the 25th percentile is likely to be within these windows: The apparent 25th percentile, plus or minus a margin related to the ε value.

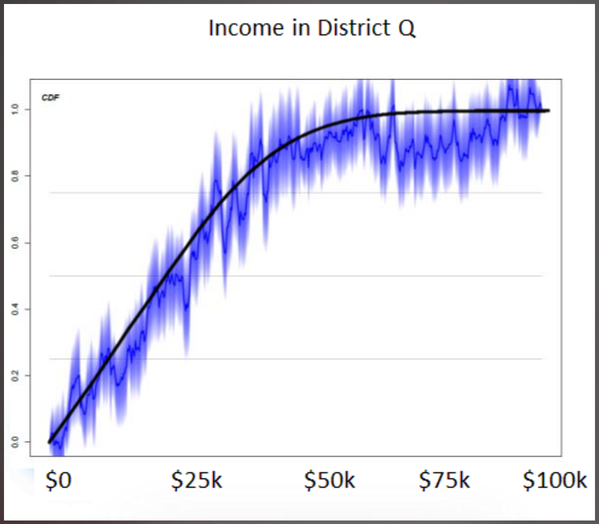
ε = 0.01 ε = 0.025 ε = 0.25



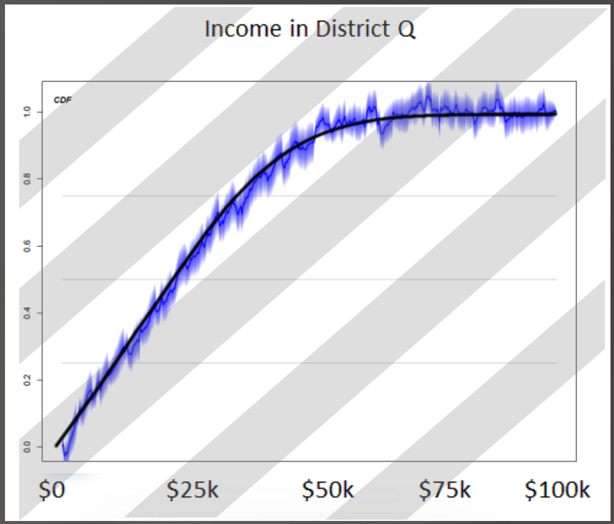
Differential Privacy in CDFs

For reference, here are the same DP-CDFs with the true CDF overlaid. We can see that the true CDF fits well within these sleeves.

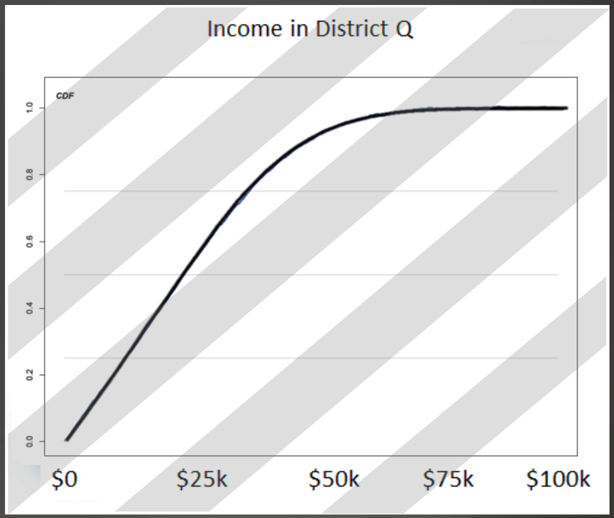
ε = 0.01



ε = 0.025

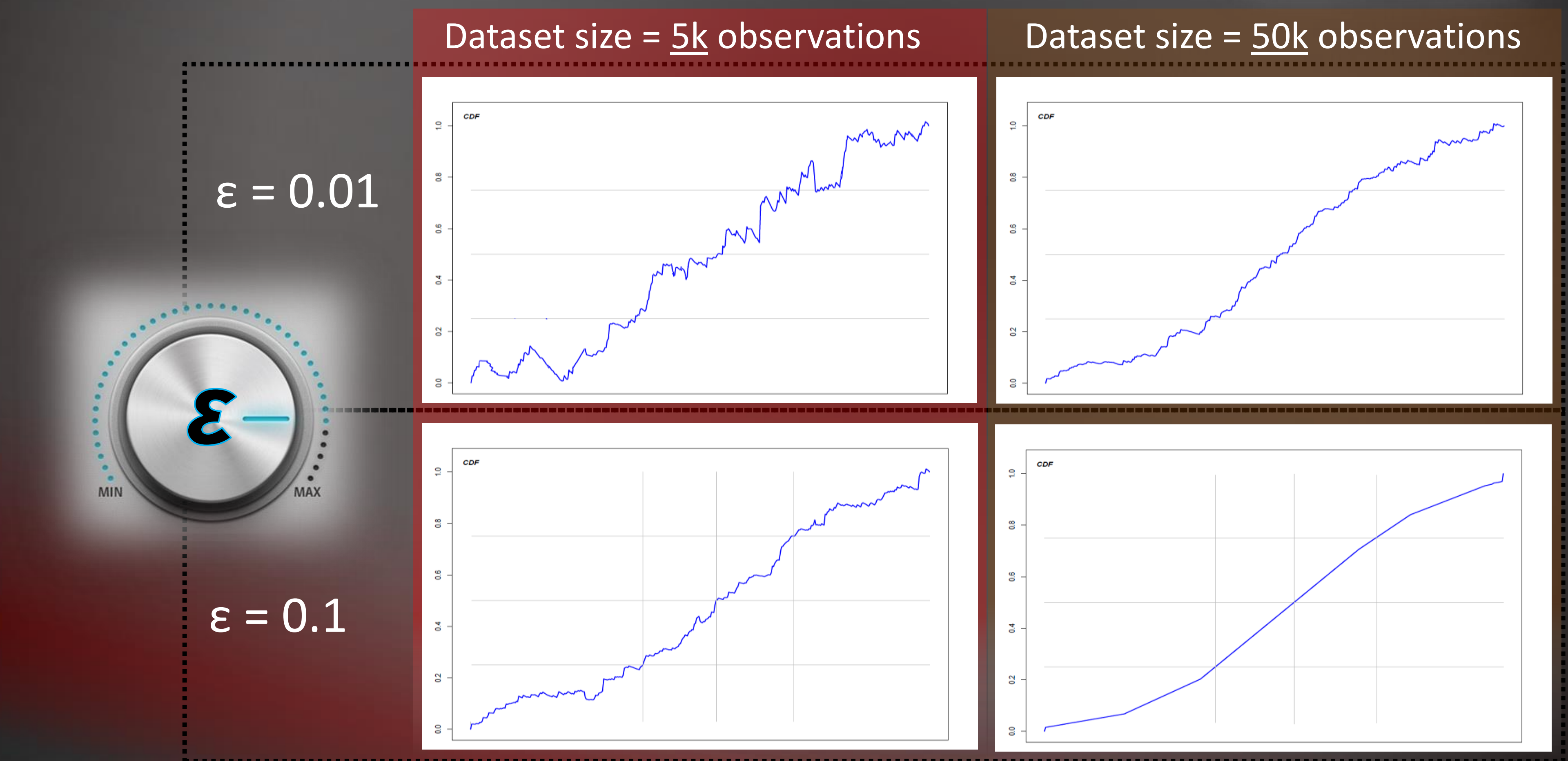


ε = 0.25



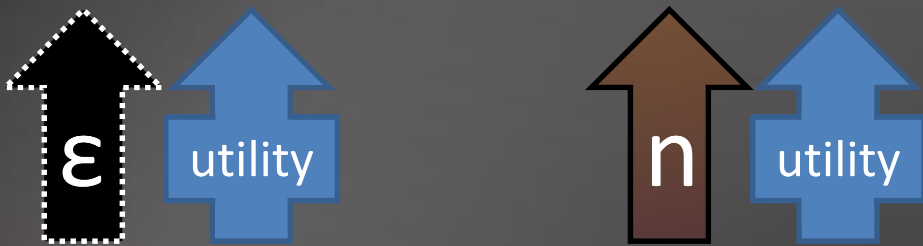
Controlling Diff. Privacy in CDFs

ε is not the only important parameter when constructing DP-CDFs. In DP-CDFs, we can also consider dataset size. The larger the set, the lower the effect of random noise. Below we have two examples of this effect.

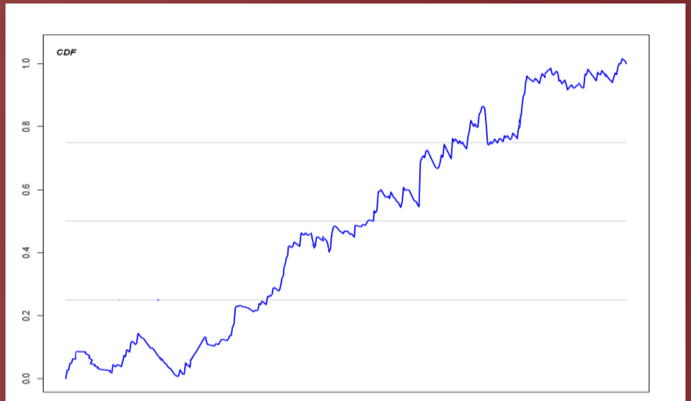
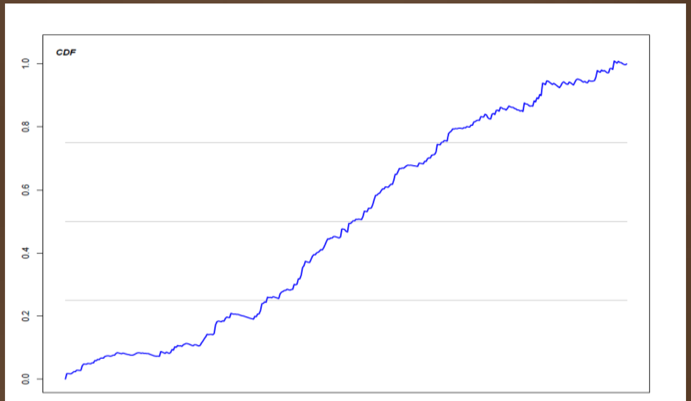
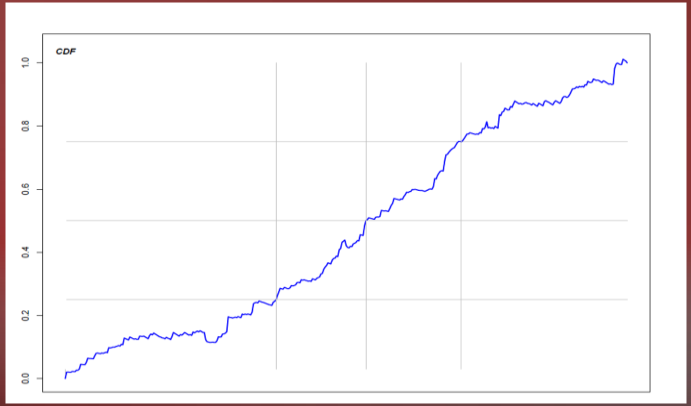
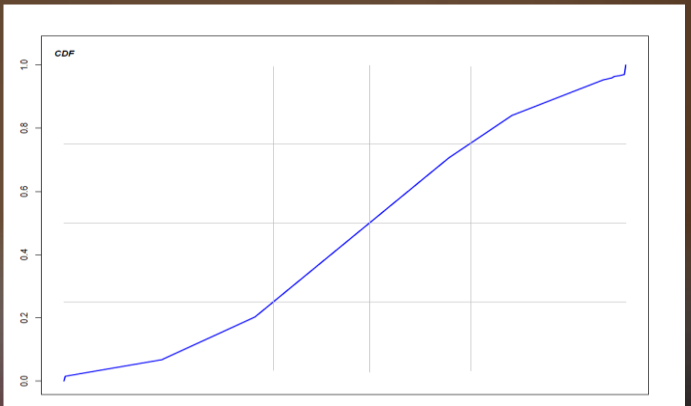


Controlling Diff. Privacy in CDFs

Thus, we have just three guidelines for interpreting DP-CDFS:



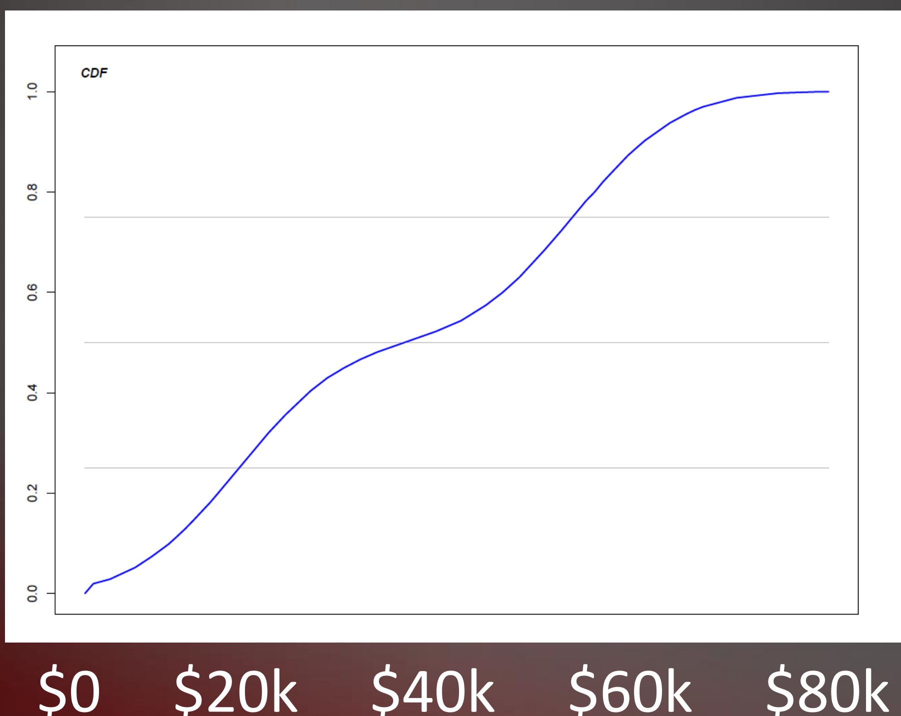
Think with sleeves

	Dataset size = 5k observations	Dataset size = 50k observations
ε = 0.01		
ε = 0.1		

Interpreting CDFs

Lastly, as an exercise here are some DP-CDFs shown without their underlying non-private CDFs. We can attempt to interpret these as we would in reality.

Income in District F



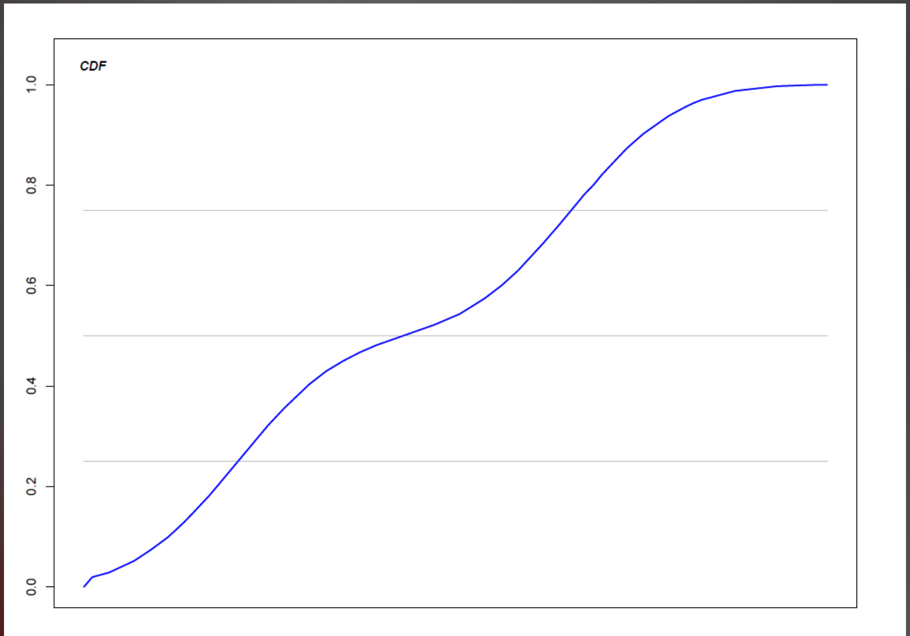
ε = .2, n = 50000

First, we have this DP-CDF with <u>high ε</u> and a <u>large dataset</u>. Because both key parameters are very high, we can assume that this approximation is fairly accurate.

Interpreting CDFs

Lastly, as an exercise here are some DP-CDFs shown without their underlying non-private CDFs. We can attempt to interpret these as we would in reality.

Income in District F



$0 $20k $40k $60k $80k

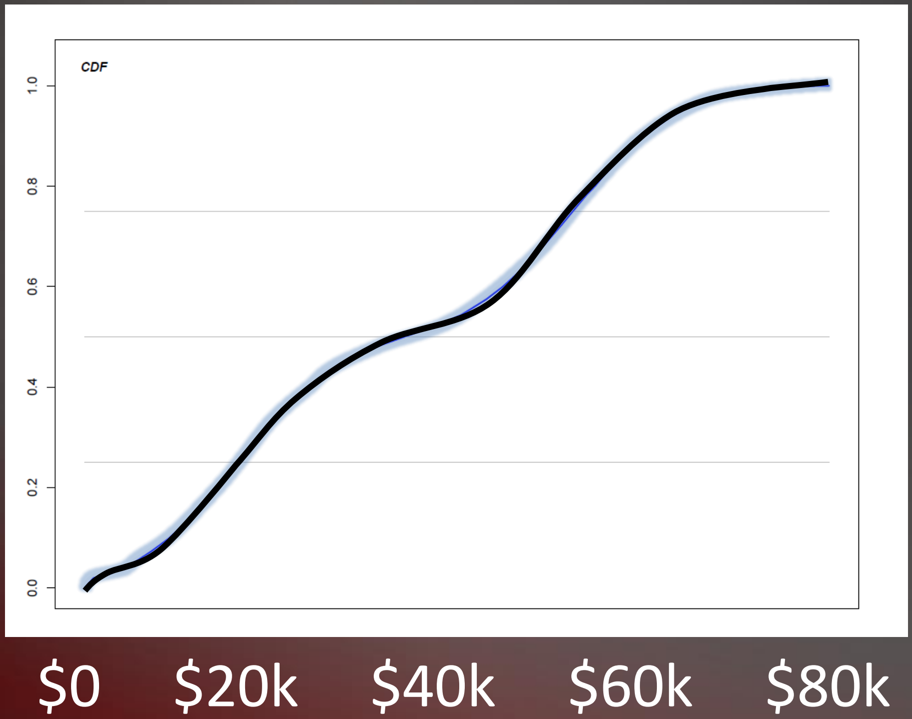
ε = .2, n = 50000

We can interpret that District F has a somewhat smooth income distribution with very few people earning more than $70k. We see two modes *around* $20k and $55k.

Interpreting CDFs

Lastly, as an exercise here are some DP-CDFs shown without their underlying non-private CDFs. We can attempt to interpret these as we would in reality.

Income in District F



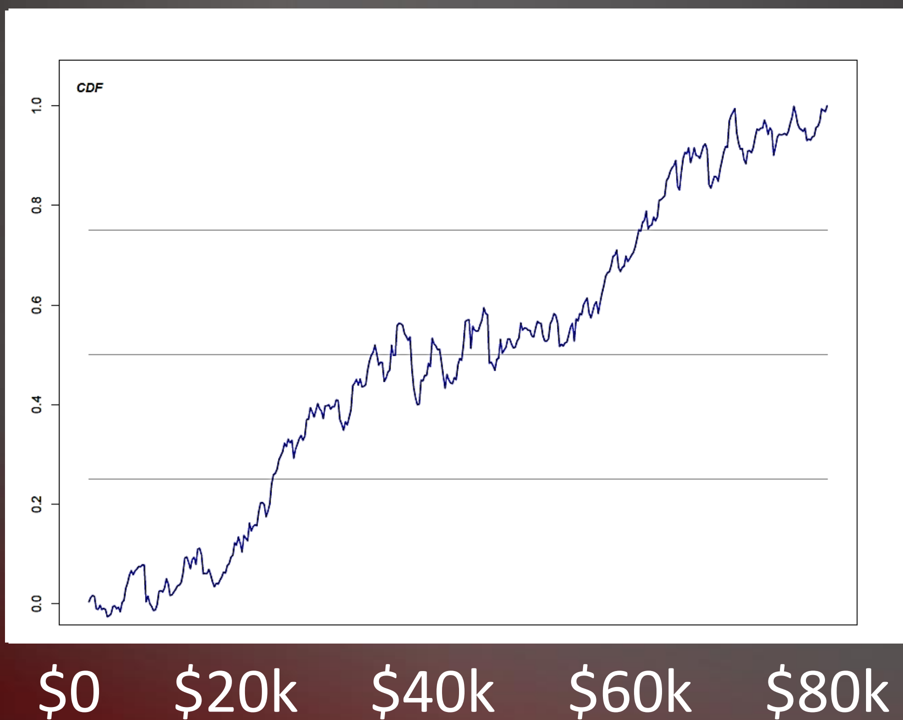
ε = .2, n = 50000

Lastly, here we've overlaid the true CDF, and we can predictably see that we have a near-perfect match. We'll now move onto less clear DP-CDFs.

Interpreting CDFs

Lastly, as an exercise here are some DP-CDFs shown without their underlying non-private CDFs. We can attempt to interpret these as we would in reality.

Income in District G



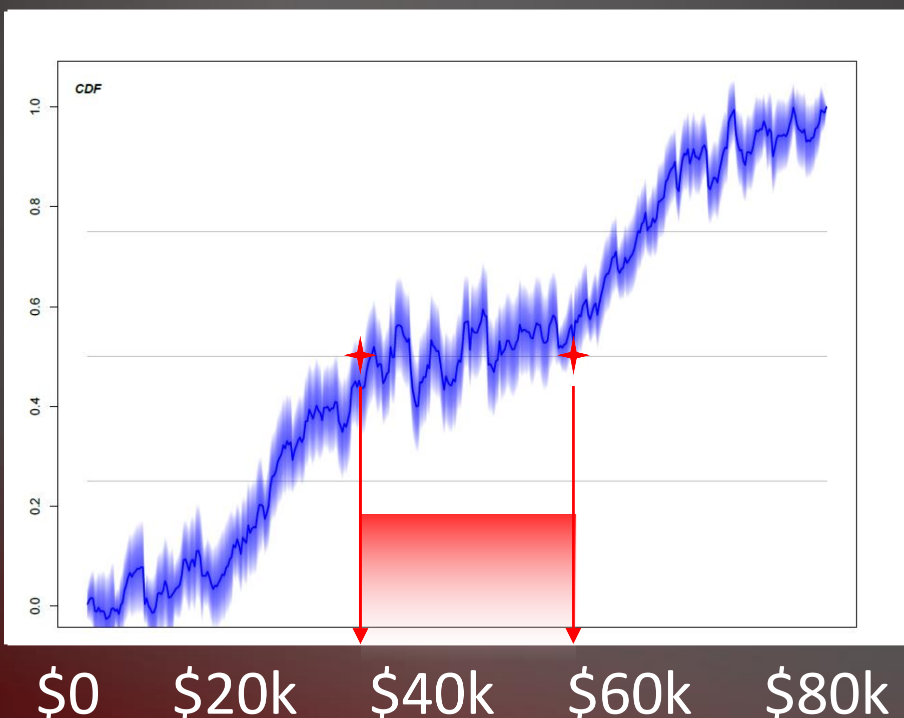
$\varepsilon = .003, n = 10000$

In this new CDF, our ε is very low. At a glance, we can infer that this income distribution is similar to the one we just saw. If we're interested in a median, we must keep in mind that there's a wide margin.

Interpreting CDFs

Lastly, as an exercise here are some DP-CDFs shown without their underlying non-private CDFs. We can attempt to interpret these as we would in reality.

Income in District G



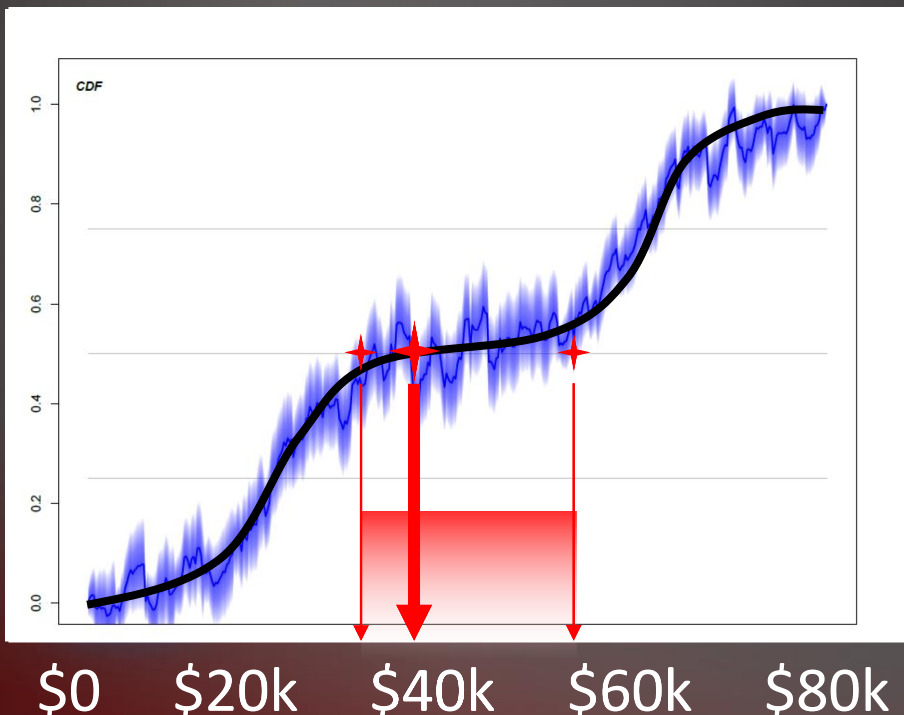
ε = .003, n = 100000

We're given a 95% confidence interval sleeve, and from that we can construct a 95% confidence interval for the median value, as shown in red

Interpreting CDFs

Lastly, as an exercise here are some DP-CDFs shown without their underlying non-private CDFs. We can attempt to interpret these as we would in reality.

Income in District G



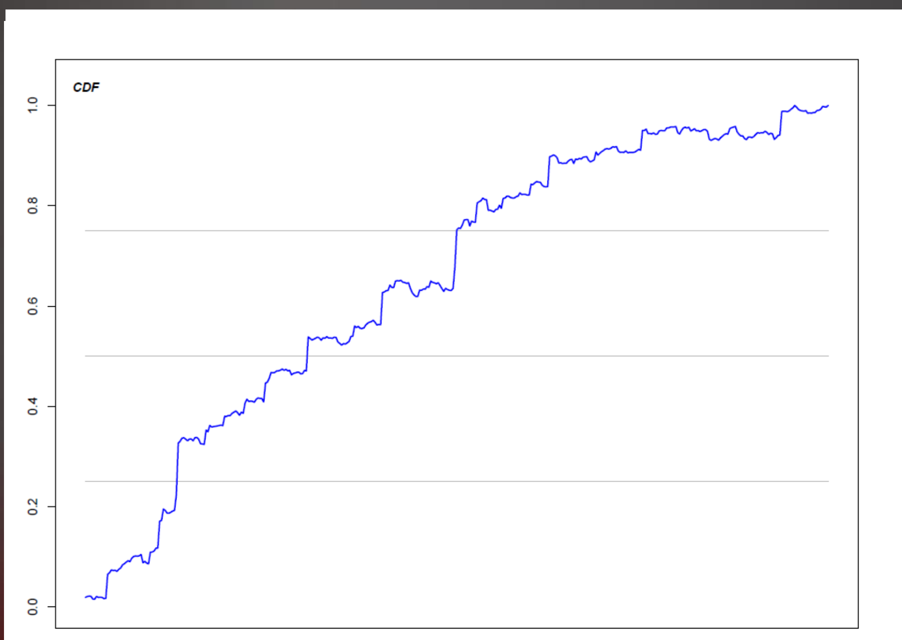
ε = .003, n = 100000

Overlaying the true CDF, we see that the median was \$38k. This illustrates that while it is tempting to interpret DP-CDFs as exact measurements, we must maintain that *true* exact points could be anywhere in the "sleeve".

Interpreting CDFs

Lastly, as an exercise here are some DP-CDFs shown without their underlying non-private CDFs. We can attempt to interpret these as we would in reality.

Income in District M



$0 $30k $60k $90k $120k

ε = .05, n = 50000

Here we see another income distribution from a wealthier neighborhood. We see sharp jumps in the curve, but it's not clear if they're from random noise or the underlying data.

Interpreting CDFs

Lastly, as an exercise here are some DP-CDFs shown without their underlying non-private CDFs. We can attempt to interpret these as we would in reality.

Income in District M

ε = .05, n = 50000

Considering our somewhat high parameter values, the 95% confidence interval is very thin. While the small jaggedness at each step could be from random noise, the step-like distribution appears to reflect the underlying data.

Interpreting CDFs

Lastly, as an exercise here are some DP-CDFs shown without their underlying non-private CDFs. We can attempt to interpret these as we would in reality.

Income in District M

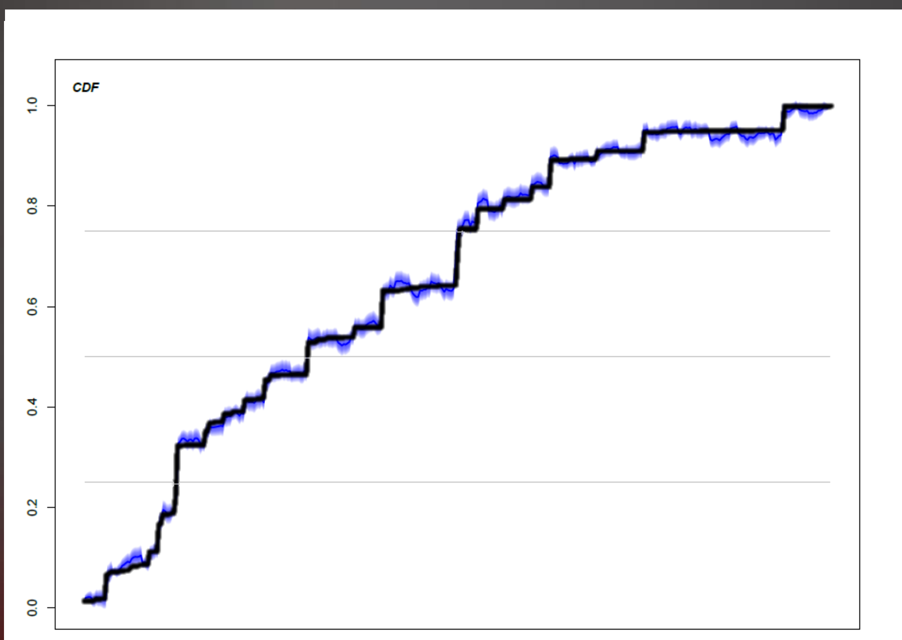
ε = .05, n = 50000

In reality, this DP-CDF closely resembled the original data. Separating the effects of noise from the effects of distribution is not always immediately clear.

Interpreting CDFs

Lastly, as an exercise here are some DP-CDFs shown without their underlying non-private CDFs. We can attempt to interpret these as we would in reality.

Income in District M



$0 $30k $60k $90k $120k

ε = .05, n = 50000

Here we have the true CDF overlaid in black. In reality, this DP-CDF closely resembled the original data.

Interpreting CDFs

Lastly, as an exercise here are some DP-CDFs shown without their underlying non-private CDFs. We can attempt to interpret these as we would in reality.

Ages in District P

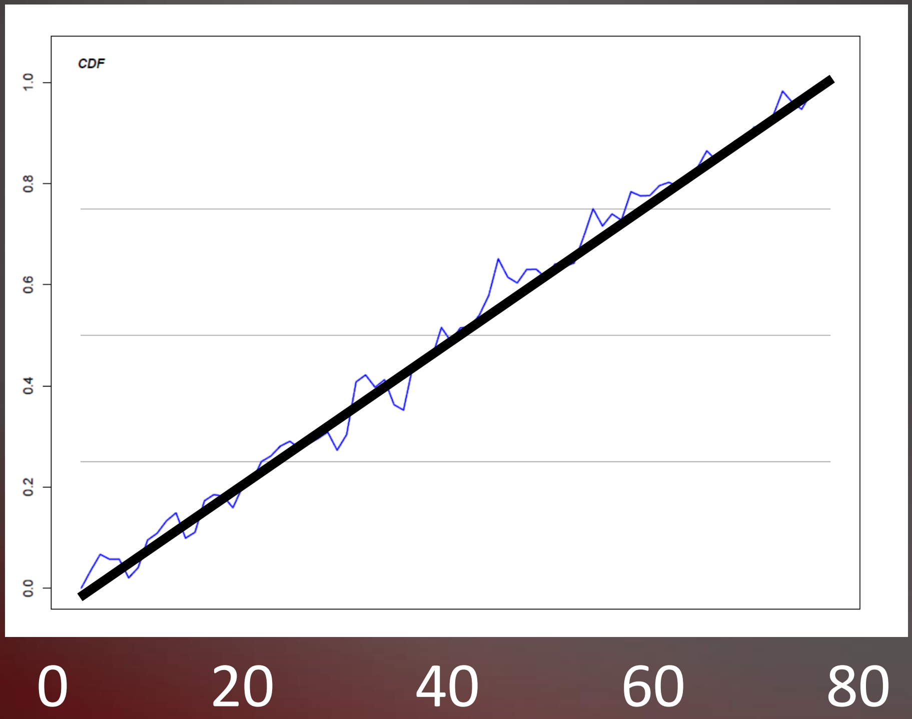
ε = .025, n = 50000

We now turn to another variable, age. This DP-CDF seems to shows a very uniform distribution of ages, but with a low ε it's not immediately clear. We are not given a confidence interval sleeve, but our best assumption is uniformity.

Interpreting CDFs

Lastly, as an exercise here are some DP-CDFs shown without their underlying non-private CDFs. We can attempt to interpret these as we would in reality.

Ages in District P (years)



$\varepsilon = .025$, $n = 50000$

Indeed, true data was shaped perfectly uniformly. We now look at one more dataset.

Interpreting CDFs

Lastly, as an exercise here are some DP-CDFs shown without their underlying non-private CDFs. We can attempt to interpret these as we would in reality.

Ages in District R (years)

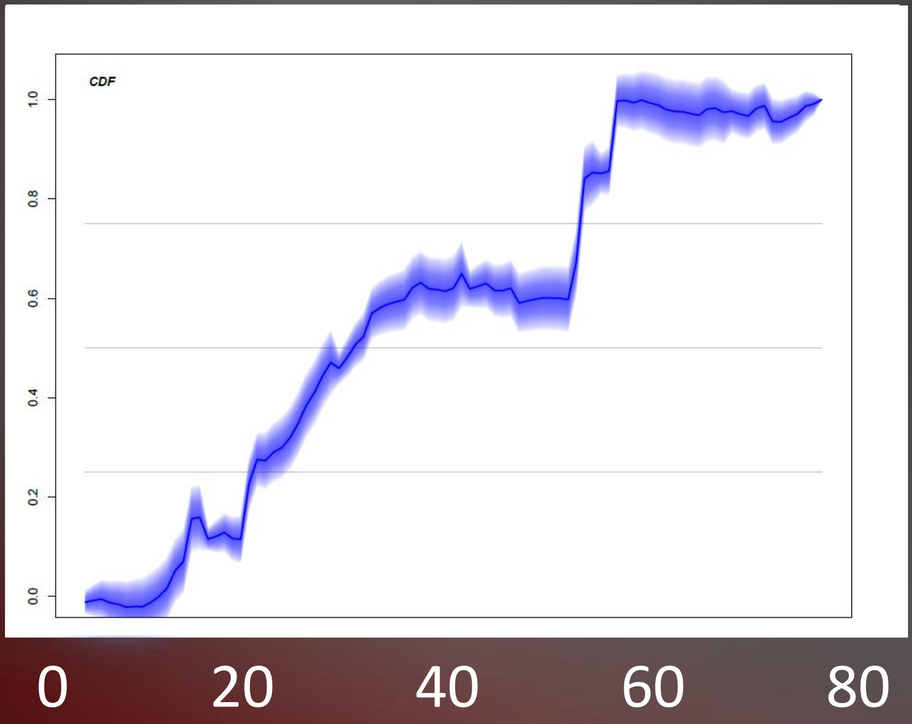
ε = .05, n = 50000

This DP-CDF is shaped abnormally. It appears to depict underlying data that is sparsely distributed. We could infer that almost nobody is aged between 30 and 50 years, but very many are aged 20 and 60 years.

Interpreting CDFs

Lastly, as an exercise here are some DP-CDFs shown without their underlying non-private CDFs. We can attempt to interpret these as we would in reality.

Ages in District R (years)



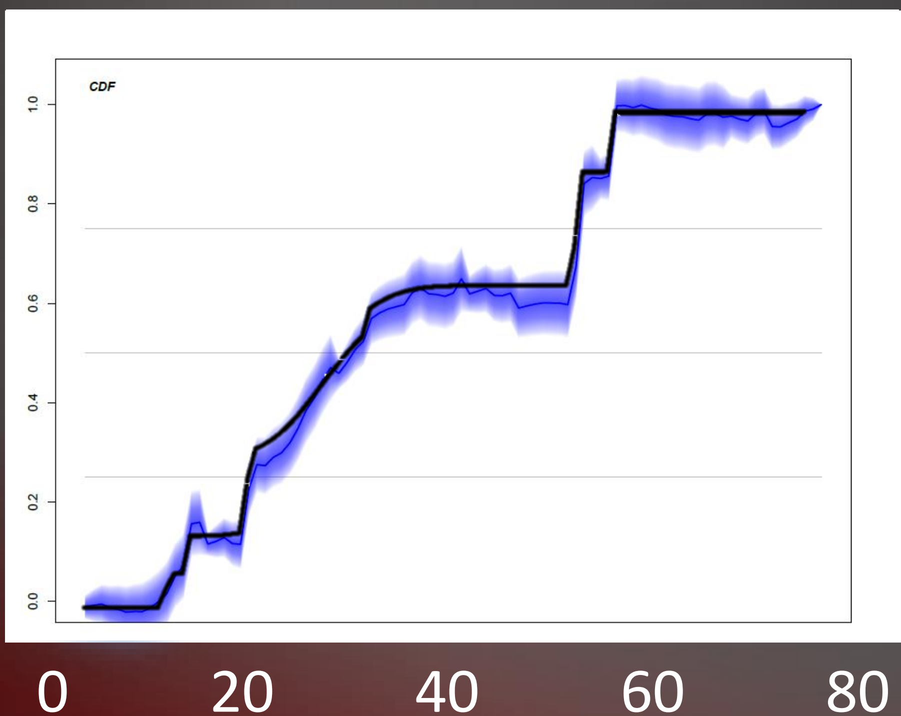
ε = .05, n = 50000

Visualizing the sleeve helps to show which features are the effect of random noise and which come from the underlying data. The most reasonable assumption is that this data is indeed sparse.

Interpreting CDFs

Lastly, as an exercise here are some DP-CDFs shown without their underlying non-private CDFs. We can attempt to interpret these as we would in reality.

Ages in District R (years)



ε = .05, n = 50000

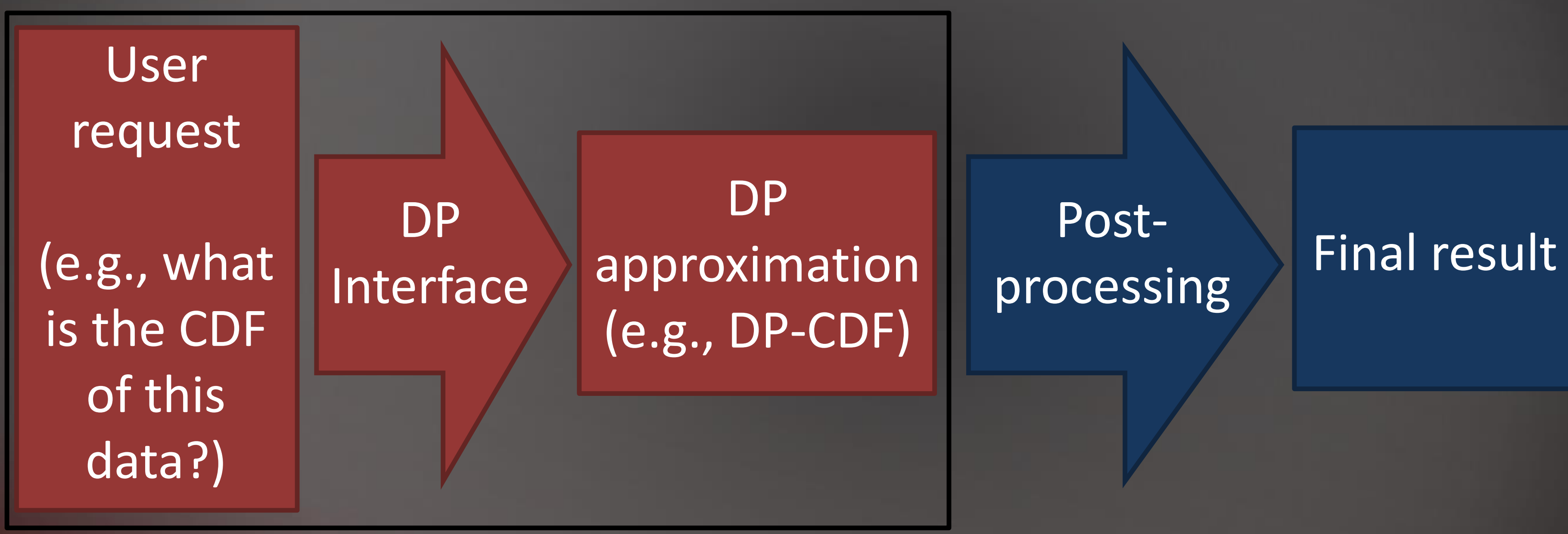
Here we see that sparse data is truly what's being represented.

Section 5

DIFFERENTIAL PRIVACY NOTES: POST-PROCESSING

Post-processing

Many applications of differential privacy incorporate an additional step known as post-processing, or manipulation to our DP-approximations that improve their utility.



DP-approximations retain the same level of privacy no matter how much post-processing they undergo.

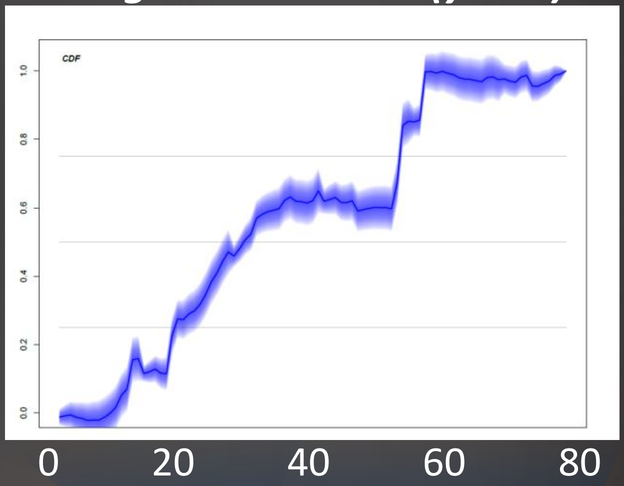
Post-processing

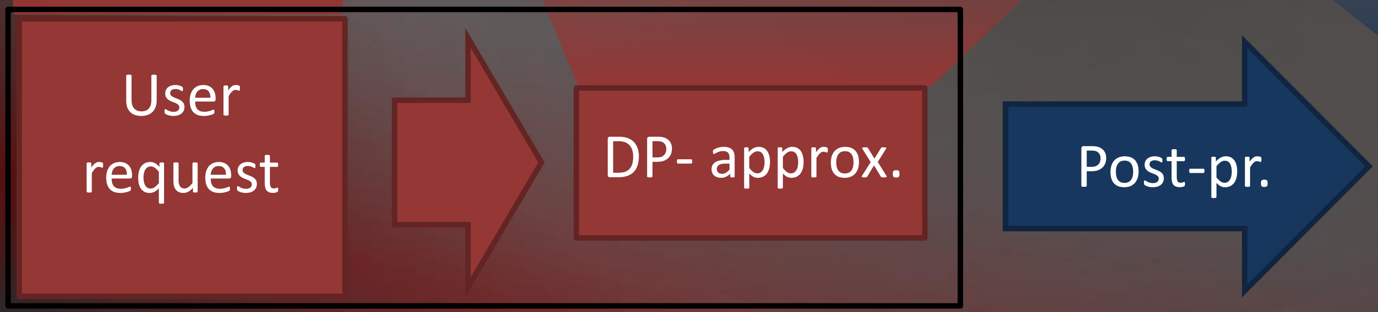
The "sleeves" we've placed around histograms and CDFs throughout this presentation are a very basic form of post-processing. They help us to interpret the DP-approximation.

CDF of **Age** in **District P**
ε = 0.5

Ages in District J (years)

Ages in District J (years)

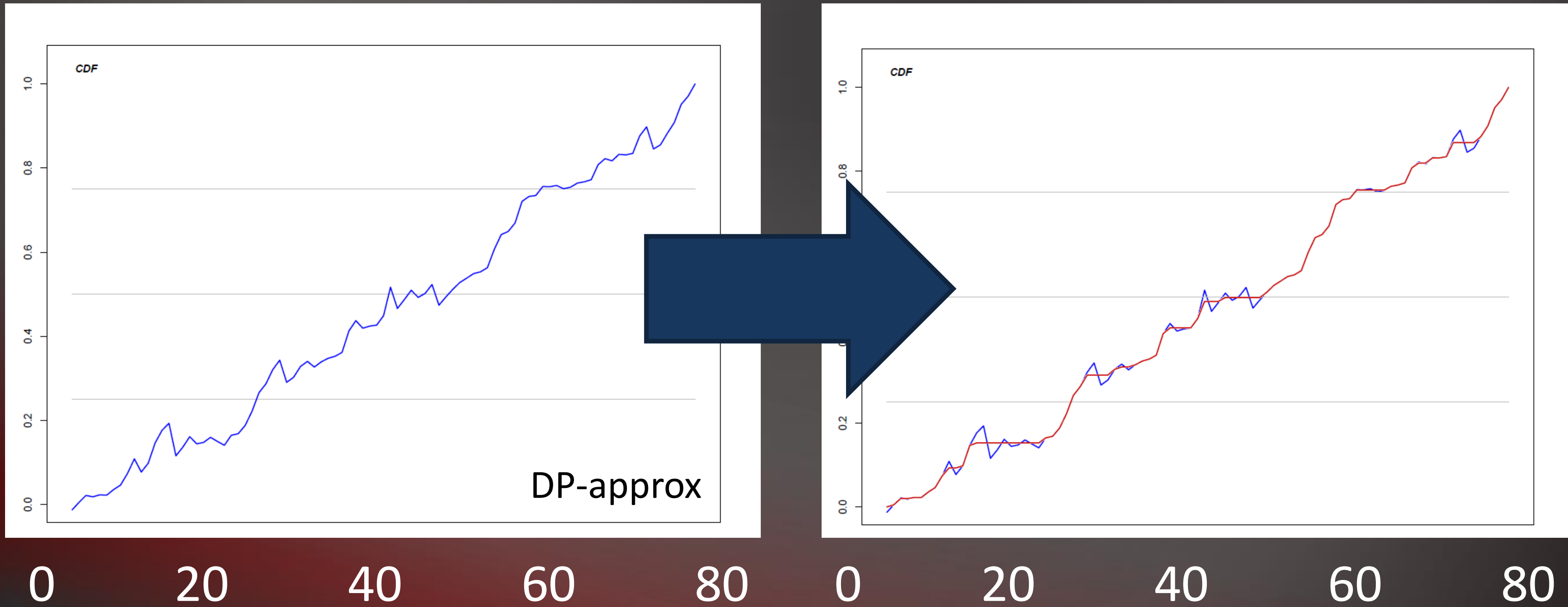




Post-processing: *Monotonization*

Another typical form of post-processing for CDFs is referred to as "*monotonization*". As the name suggests, this is the enforcement of monotonicity, or ensuring that no part of the CDF dips downward.

Ages in District P ***Ages in District P***



Post-processing: *Monotonization*

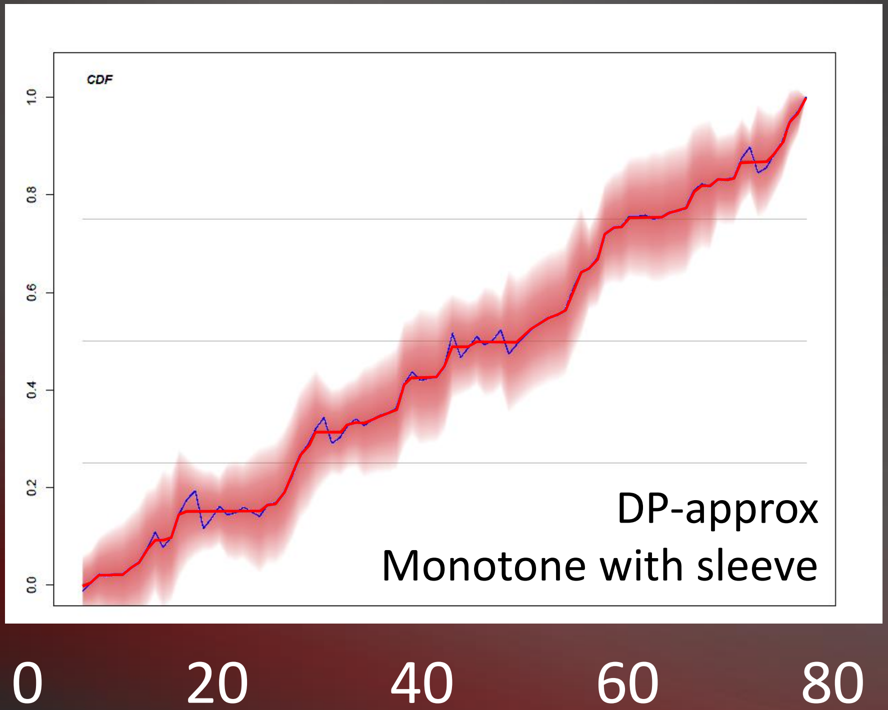
Since a CDF represents cumulative probability, negative slope implies negative probability, which is not possible. Any negative slope in a CDF is the product of the random noise of differential privacy. Monotonization resolves this. However, monotonization also introduces vertical spikes and horizontal plateaus where they may not have been previously. These visual effects of differential privacy are known as "artificial artifacts".

Ages in District P ***Ages in District P***

Post-processing: *Monotonization*

For that reason, researchers utilizing monotonization post-processing should be aware that vertical lines in the CDF do not necessarily mean that the data is tightly clustered at those values, nor do horizontal lines always suggest no data.

Ages in District P



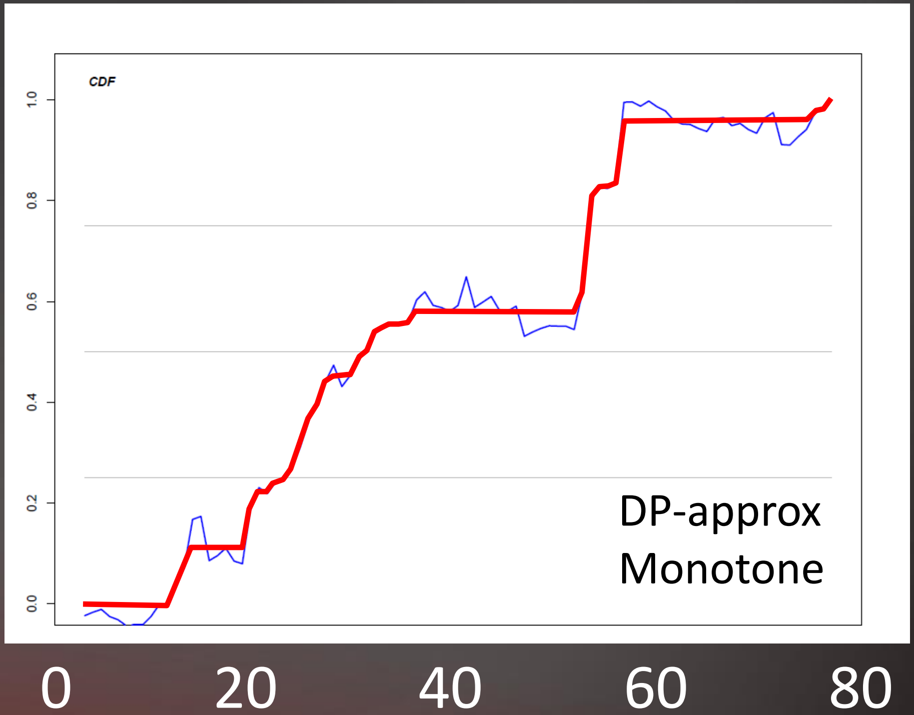
Consider that the ages in District P are actually *uniformly* distributed. Reading a monotonized CDF too literally would be a mistake. Here, visualizing the "sleeve" is most useful.

Post-processing: *Monotonization*

In cases where the real CDF is sparse: it has vertical and horizontal jumps, monotonization is not the only source for this jagged lines.

Ages in District P

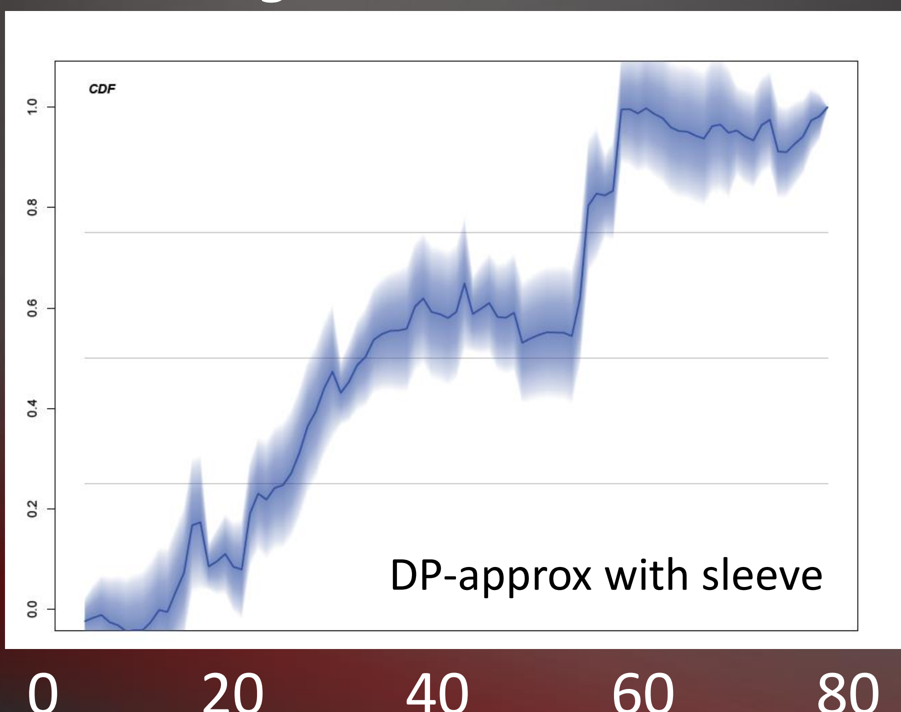
Ages in District P



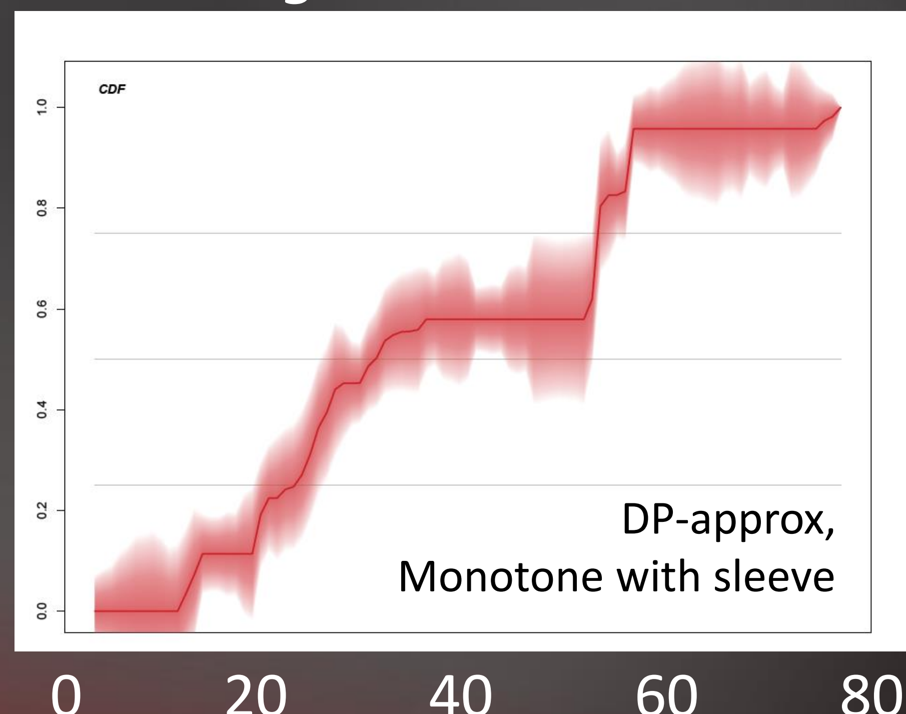
Post-processing: *Monotonization*

By using a visual sleeve on our monotonized CDF, we can imagine that jagged corners within the sleeve are somewhat likely to be the result of noise, while larger directional changes represent the underlying CDF.

Ages in District P



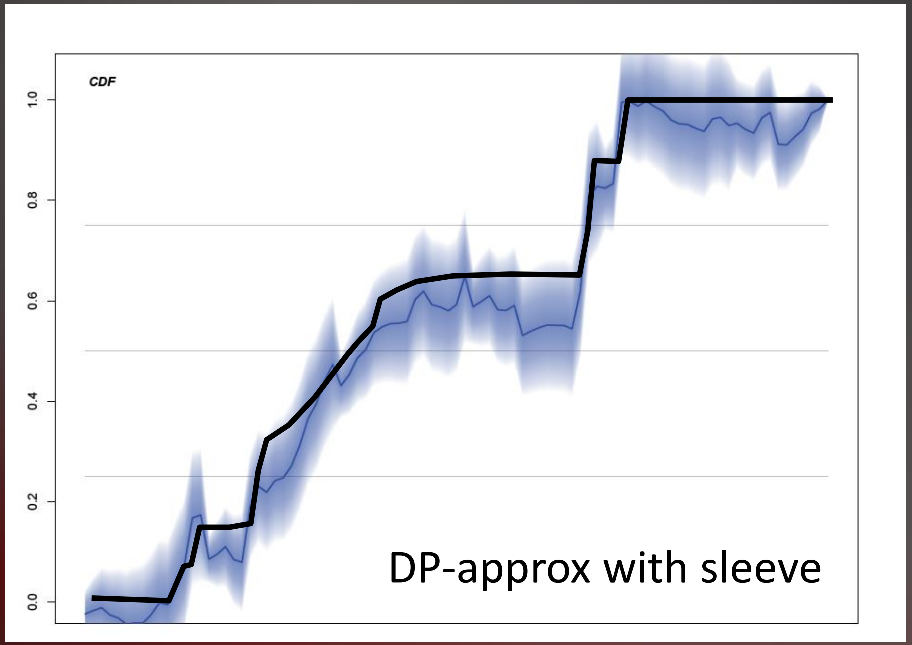
Ages in District P



Post-processing: *Monotonization*

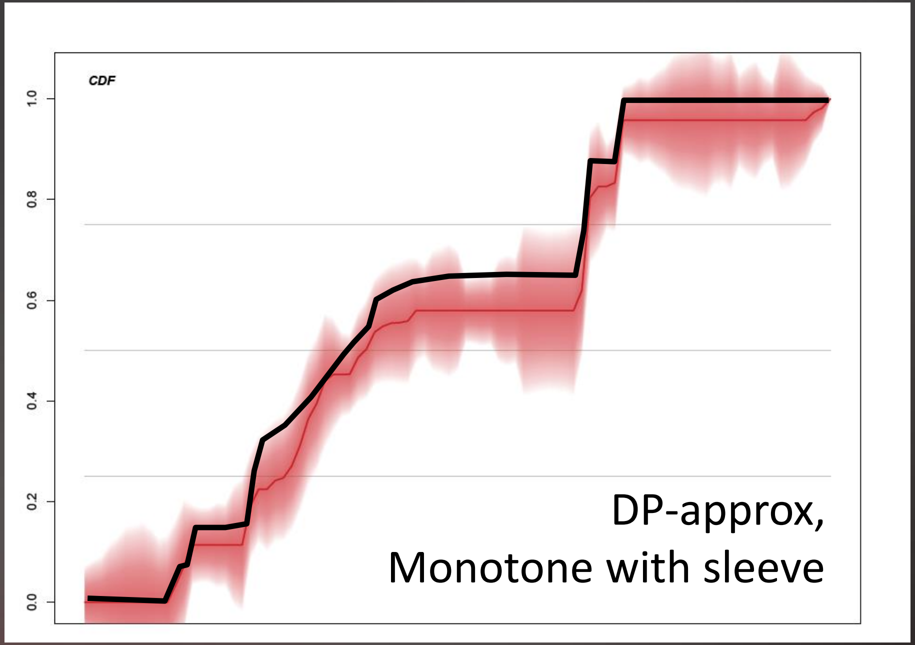
Indeed, by overlaying the true CDF, we see that the smaller jagged lines and spikes are not present, but larger vertical and horizontal segments and corners are. Either way, the 95% confidence interval sleeves appear to hold.

Ages in District P



0 20 40 60 80

Ages in District P



0 20 40 60 80

Privacy Tools
for Sharing Research
Data

*A National Science Foundation
Secure and Trustworthy Cyberspace
Project*

*with additional support from the Sloan Foundation and
Google, Inc.*

Differential Privacy in CDFs

For documents on theory, law, and other DP-statistics usage, see:

http://privacytools.seas.harvard.edu/

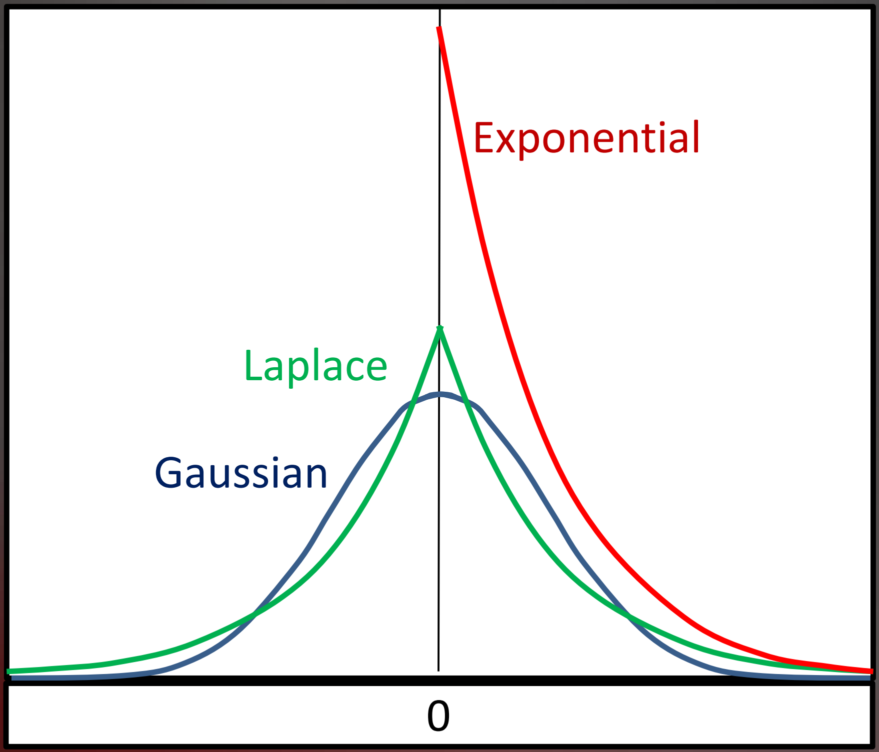
Section 6

APPENDIX: UNDEVELOPED

[Different Algorithms]

Differential Privacy: Noise

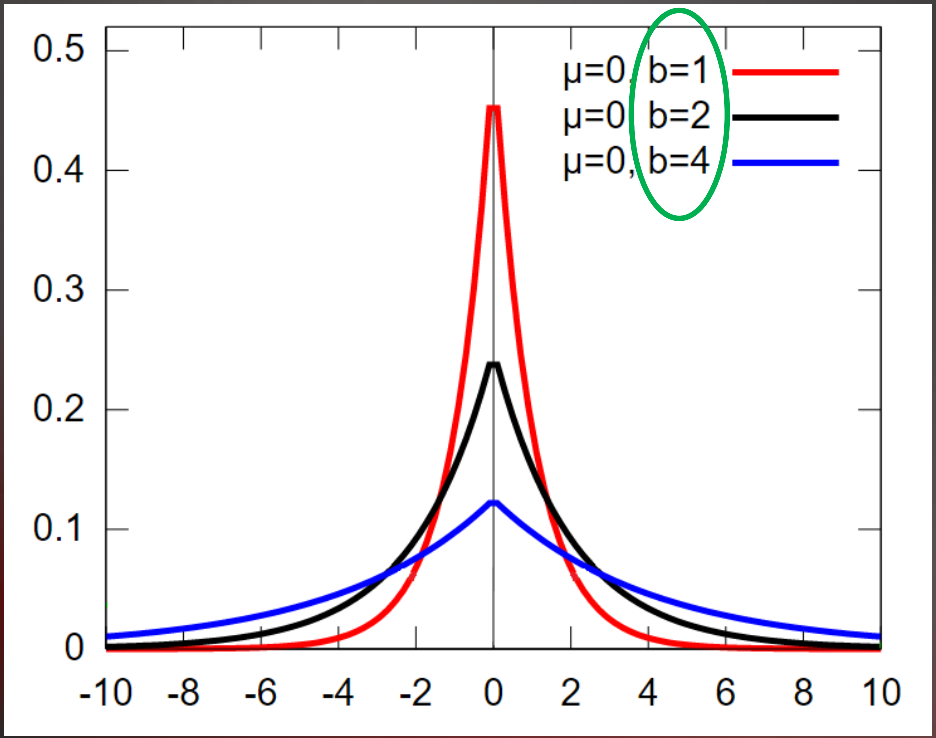
Other documents from this group provide mathematical definitions for the inner-workings of differential privacy. This document will simply cover the intuition of differential privacy as far as is useful for statistical work.



Differential privacy is based on the introduction of random noise, which relies on the usage of a random variable. Random variables can follow several distributions, as depicted to the left.

Differential Privacy: Noise

Below are three representations of Laplace distributions, which are a common choice of distribution in DP mechanisms. These three curves depict the likelihood of a random variable (following each distribution) to take on a given value.

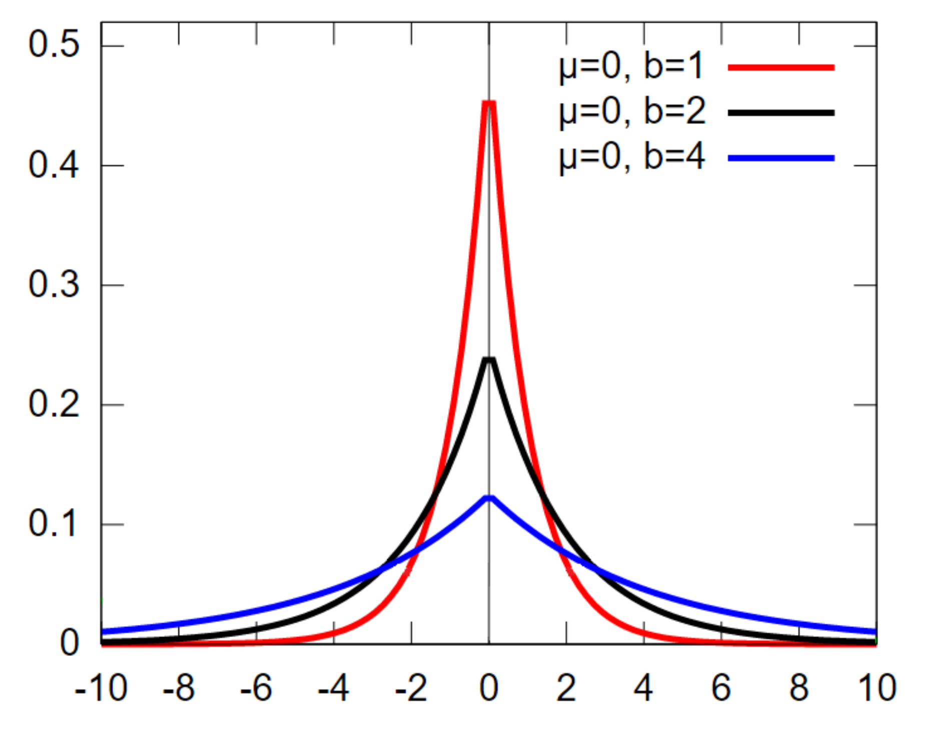


The tightness of each distribution around its mean (0 here) is controlled by the variance parameter "$b = \frac{1}{\varepsilon}$".

The Gaussian and exponential families of distributions are controllable by epsilon in a similar way.

Differential Privacy: Noise

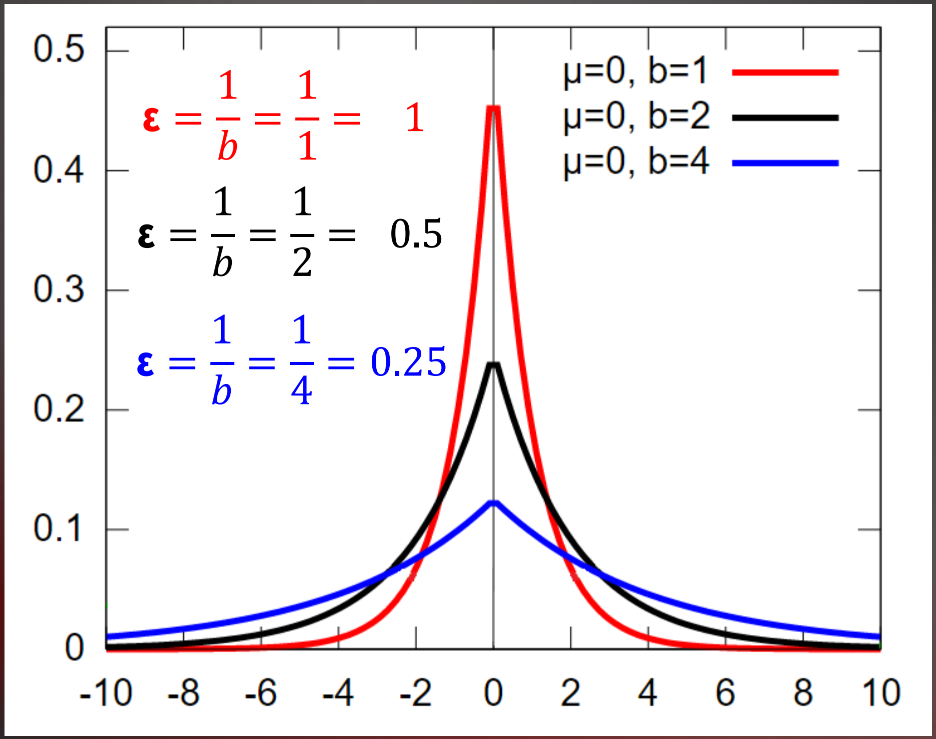
To generate "random noise", differentially private Laplace mechanisms randomly select from a certain Laplace distribution. This random selection is added onto the desired data point or statistic, thereby masking that point or statistic's true value. In most computations, μ = 0, and b is varied.



Notice that as "b" changes, the density of the distribution around 0 changes. It is up the user to define b, thus defining the likelihood of greater absolute values of noise.

Differential Privacy: ε

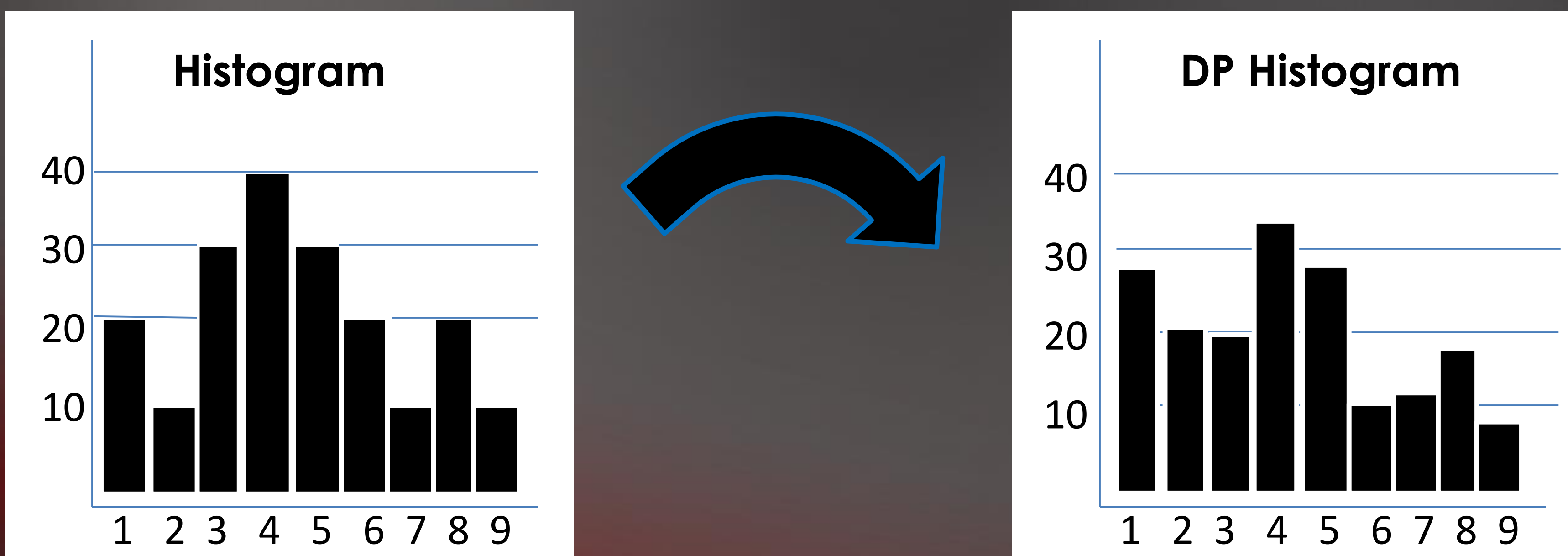
In formal definitions and in practice, differentially privacy manages ***b*** through a parameter called **ε** (epsilon), which is equal to the inverse of b.



With smaller **ε**, the possible values for random noise are more spread out away from zero. Thus, a small **ε** masks data better, and we say that data masked by Laplace (D.P.) mechanisms using smaller **ε** values are "more private" than those using larger **ε** values.

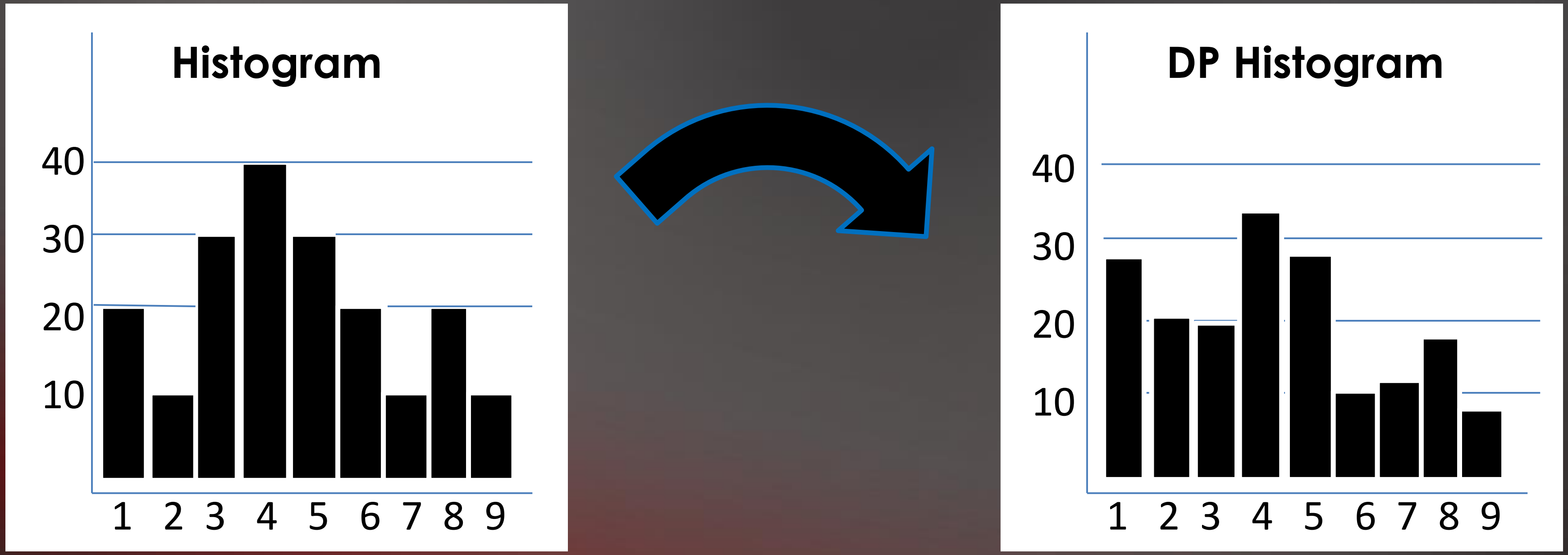
Differential Privacy: Histograms

Consider the histogram. The simplest way to create a differentially private histogram is to generate the histogram and treat the height of each bin as a single data point. Add random noise to those bin-heights to mask their true values.



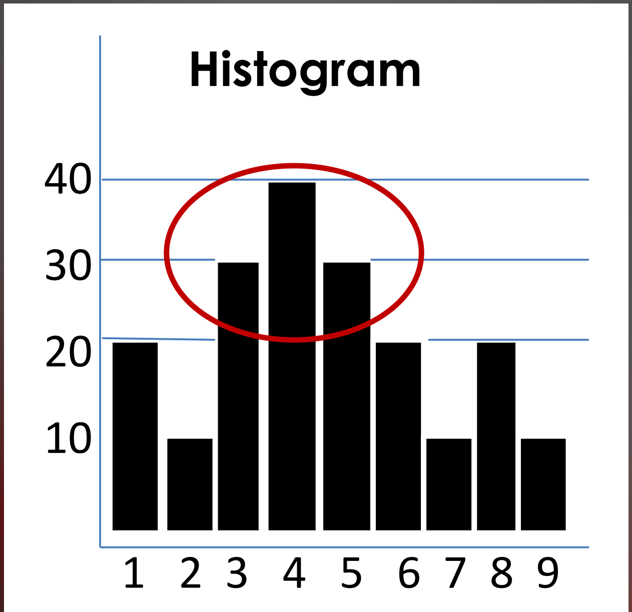
Differential Privacy: Histograms

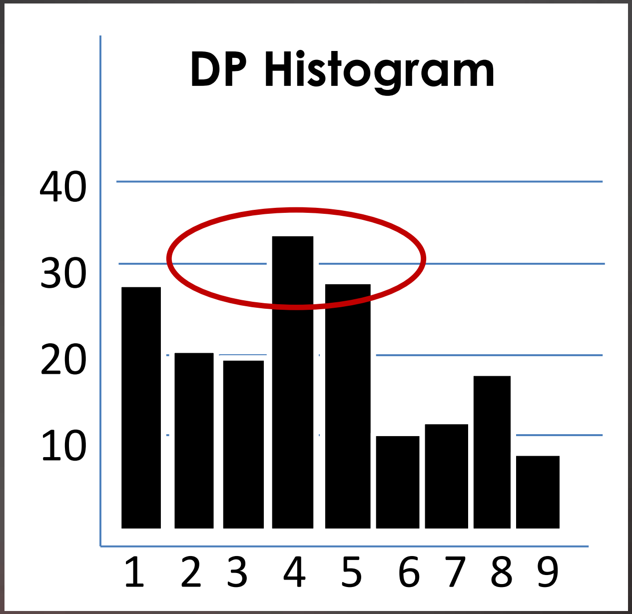
With differential privacy noise, we are in control of when and how random noise is introduced to our statistics. We can use this to our advantage by inspecting precisely what random noise has done to our histogram.



Differential Privacy: Histograms

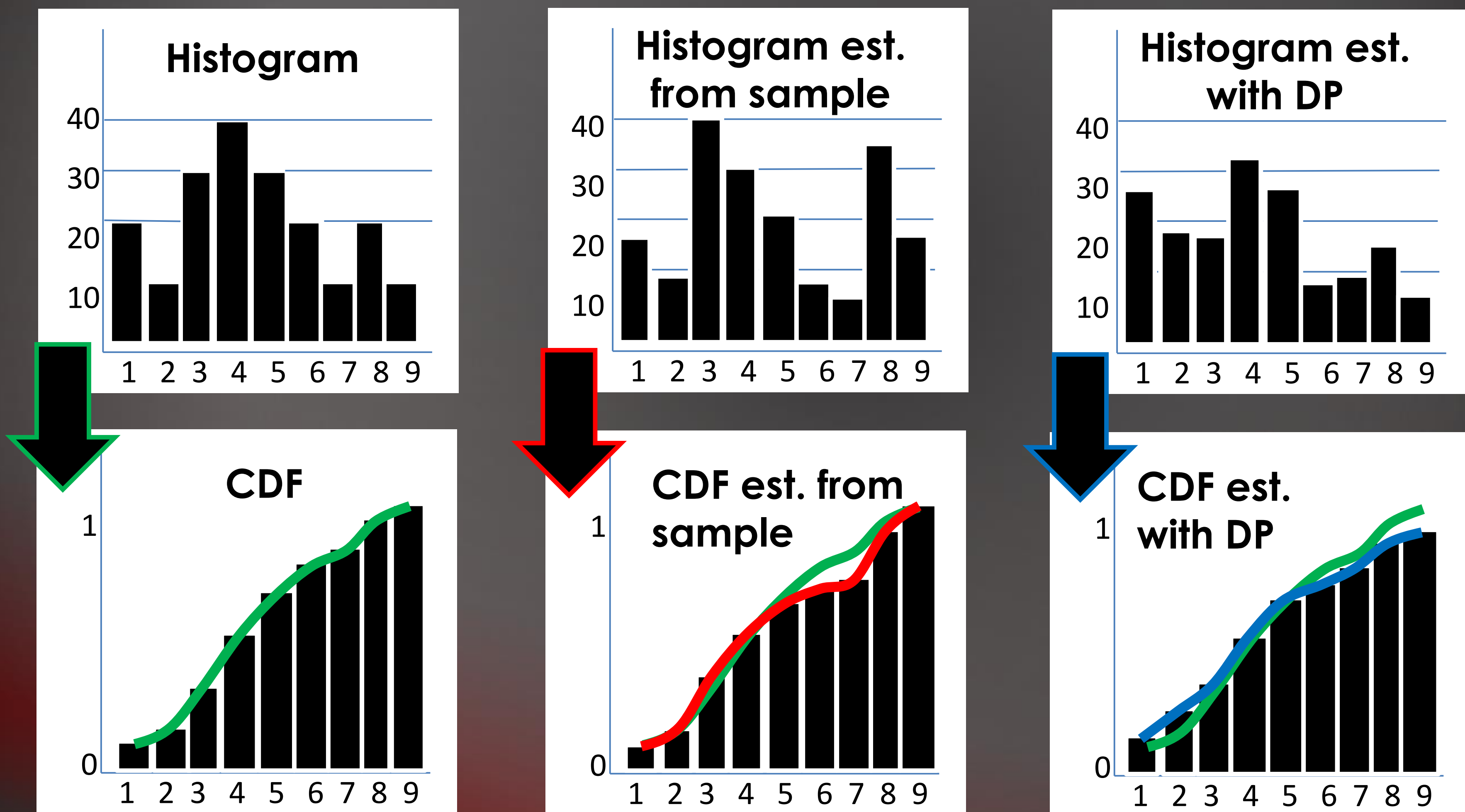
Looking closer, this transformation has strong implications. Differential privacy has the potential to change which value is shown as the mode. While looking at this DP Histogram, we can know the *distribution of possible heights* for each bin. However, if our goal is to find the mode, we need to keep in mind that it may be 3,4, or 5 (or any other value, but with somewhat negligible probability.)





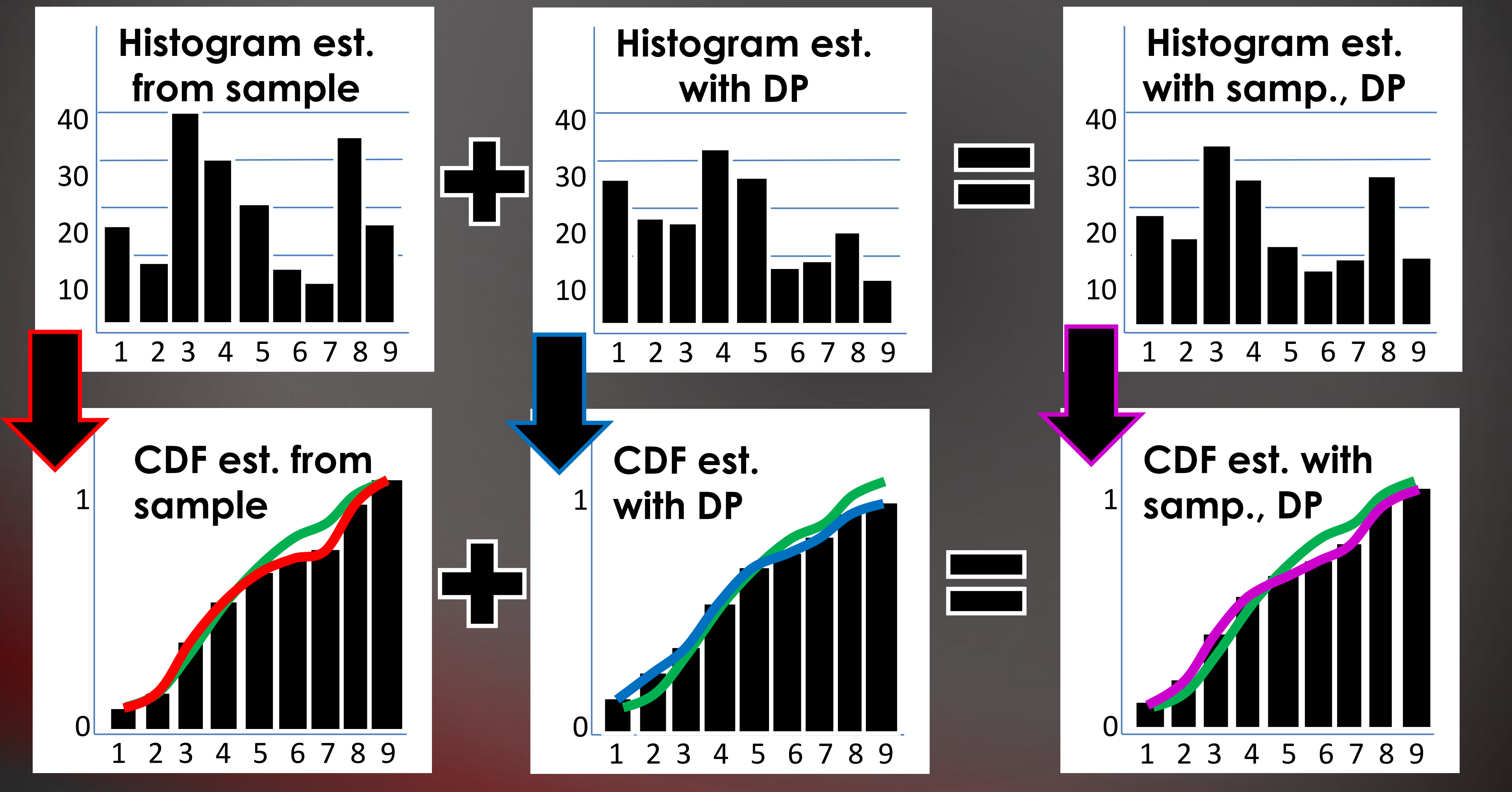
Differential Privacy: CDFs

Unlike with sampling error, DP-noise accumulates as DP-histogram bins are added. In practice, the final bin of a DP-CDF is manually set to 1, but here we can see the result of compiled variance. MOVE THIS STUFF TO THE NEXT SECTION, BE GENTLER ABOUT IT. ALSO MENTION THAT THIS IS A SIMPLE WAY FOR EXAMPLE. OUR ALGORITHMS ARE MORE SOPHISTICATED TO TRY AND REDUCE THE NOISE



Differential Privacy: CDFs

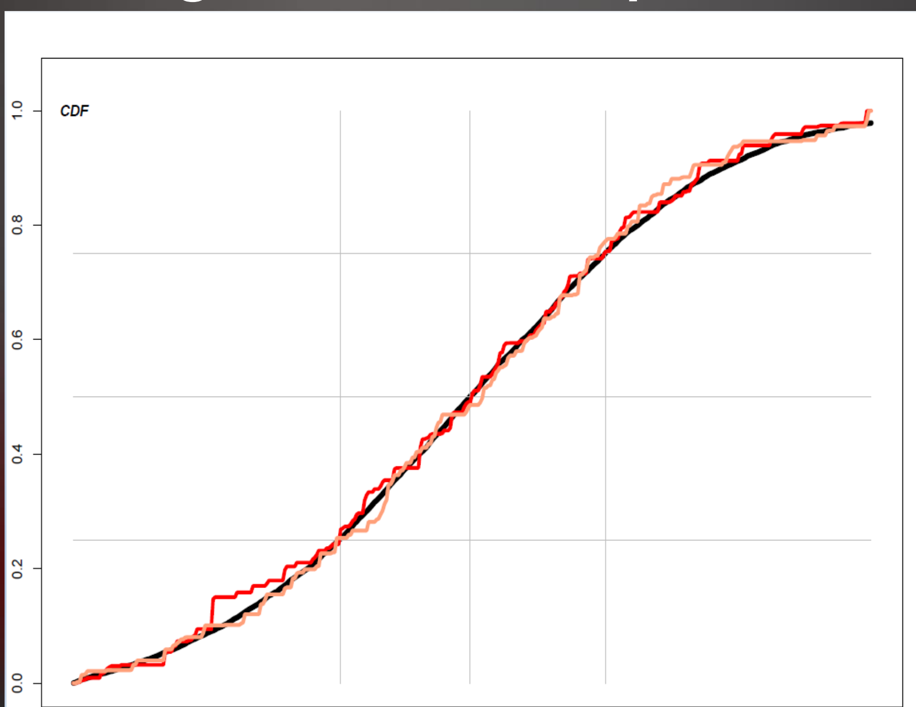
We'll finish by pointing out that DP (like most statistics) is mainly used on sampled datasets, so DP and sampling error will be simultaneously present.



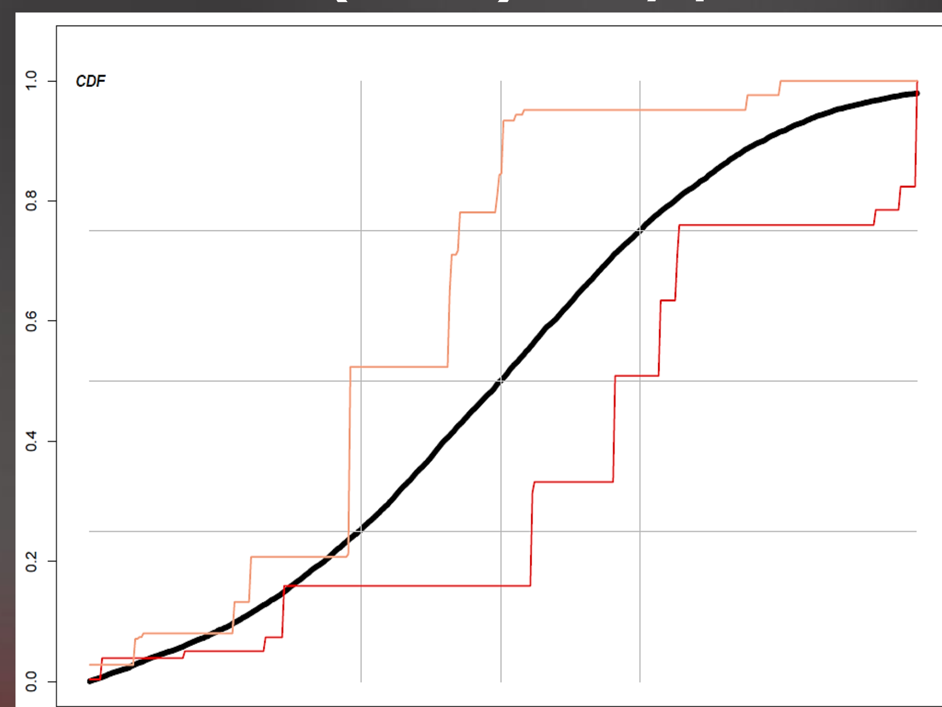
Controlling Diff. Privacy in CDFs

When we change our ε values to something lower (more private) we find that our dp-CDFs do a better job of masking the true data, and a worse job of conveying accurate information.

Large ε, 0.300, less private



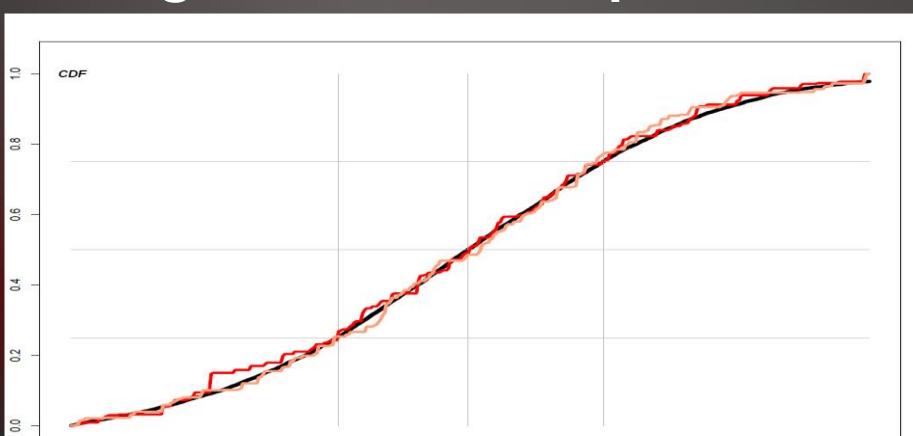
Small ε, (0.005), very private



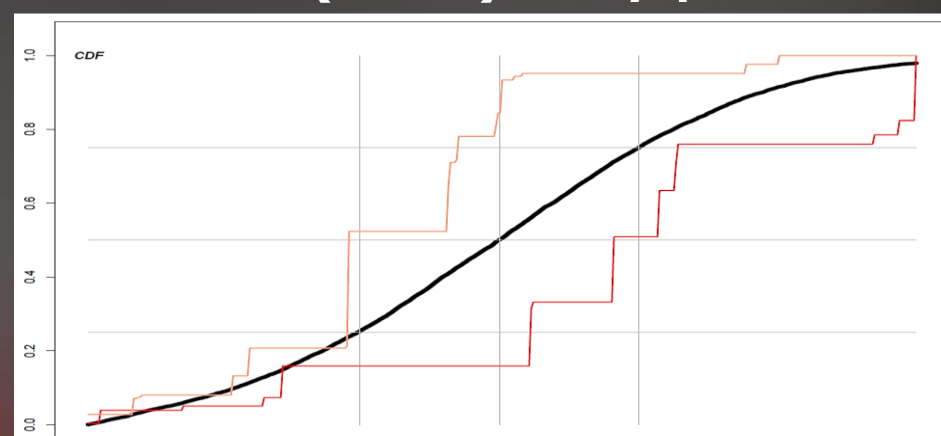
Controlling Diff. Privacy in CDFs

Consider the dp-CDFs below. While the CDFs on the left are very accurate and useful, they cost 0.3 ε each. Meanwhile, one user can "afford" many more of the less useful, more private dp-CDFs on the right.

Large ε, 0.300, less private



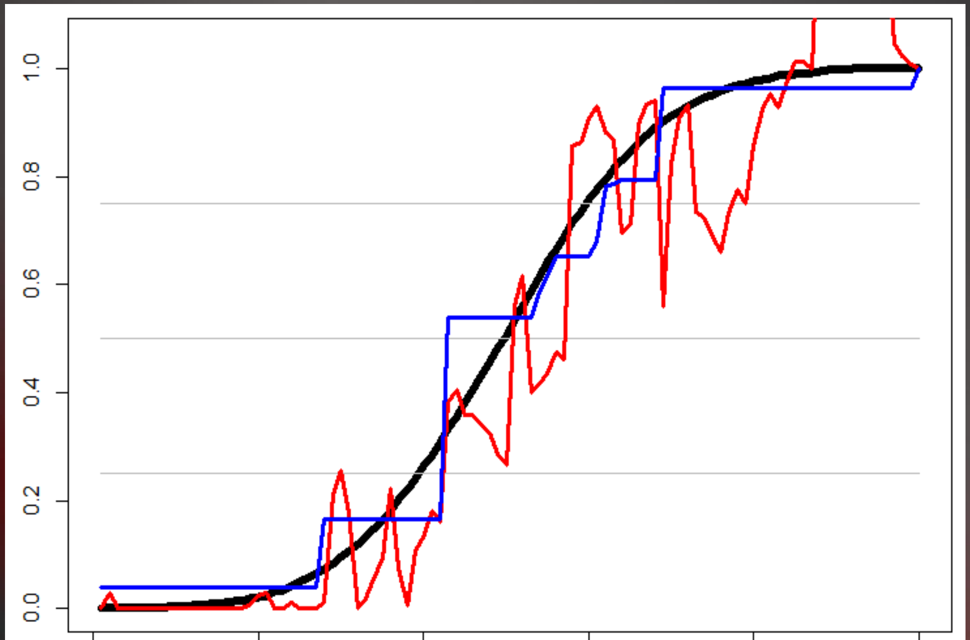
Small ε, (0.005), very private



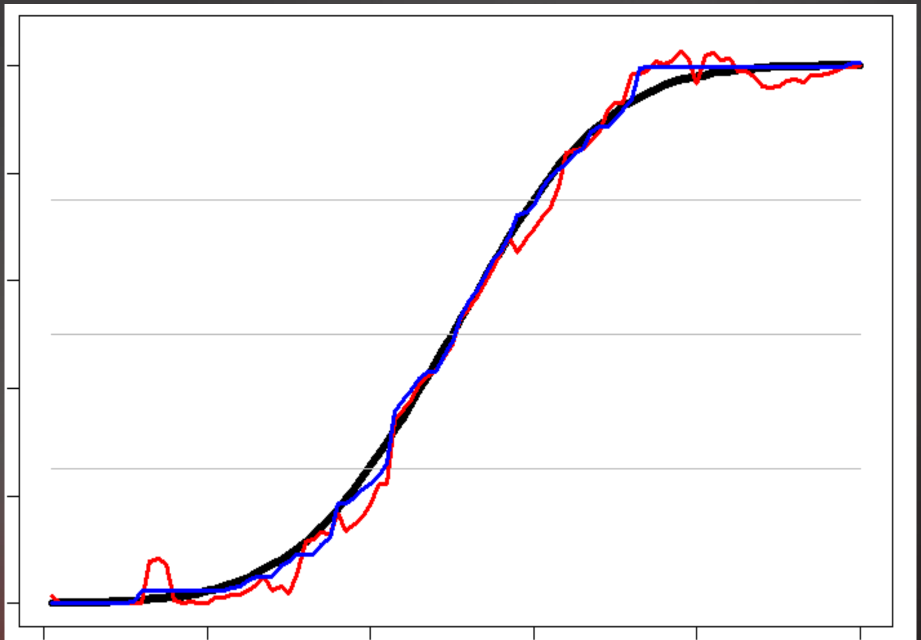
Controlling Diff. Privacy in CDFs

Under those conditions, it's clearer why we'd like more "efficient" dp-CDFs at the lowest ε price. Consider the graphs below for an example of how different methods can be more (or less) cost-effective than others. RESEARCHERS CAN A COLLECTION OF CDF ALGS. NOT ALL ARE EQUAL. EVEN WITH THE SAME EPS, SOME WILL DO BETTER OR WORSE, SOMETIMES DEPENDS ON OTHER PARAMS. HERE ARE 2 ALGS.

ε = 0.01



ε = 0.1

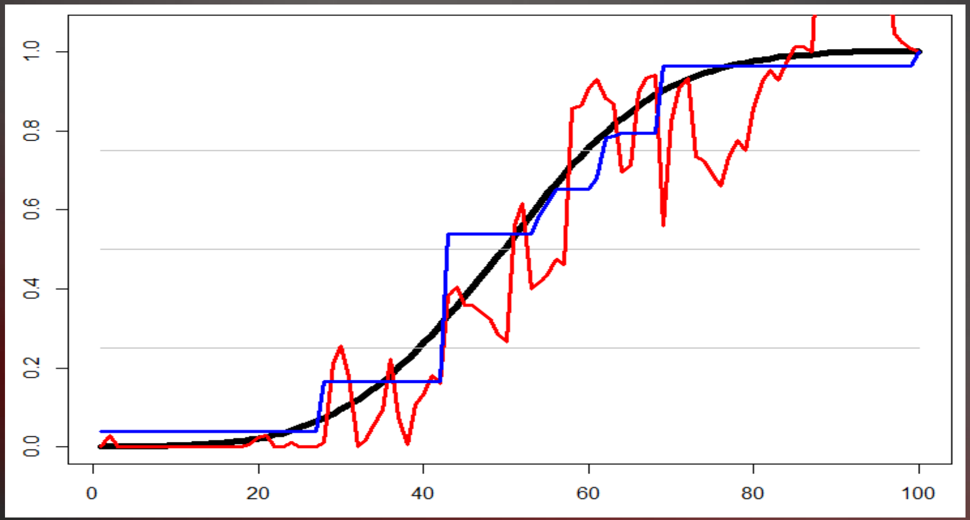


Controlling Diff. Privacy in CDFs

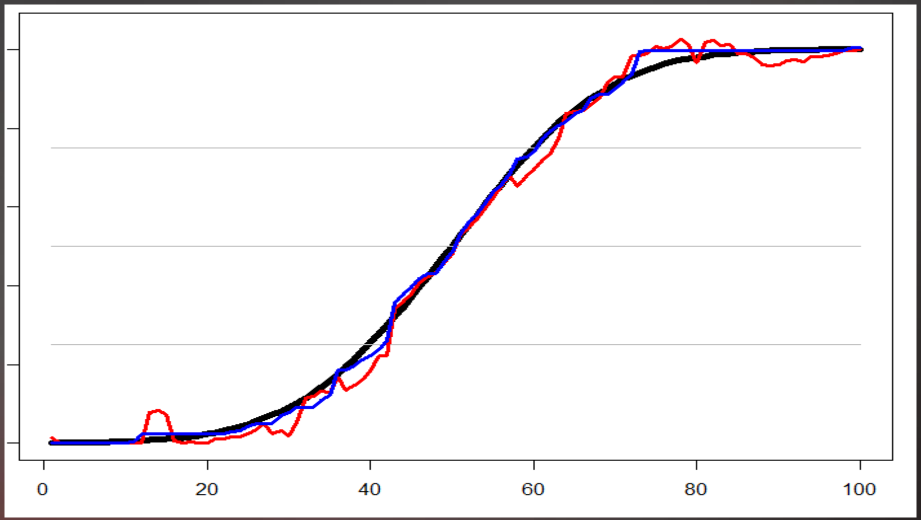
In this case, the blue curves are processed versions of the red curves, providing more useful dp-CDFs for the exact same ε price. YO SLOW DOWN. WE'RE NOT TALKING ABOUT COMPOSITION OR POST-PROCESS. APPENDIXS IF DESIRED

Ongoing improvements like this suggest that dp-stats may become yet more cost-effective in the future.

ε = 0.01



ε = 0.1

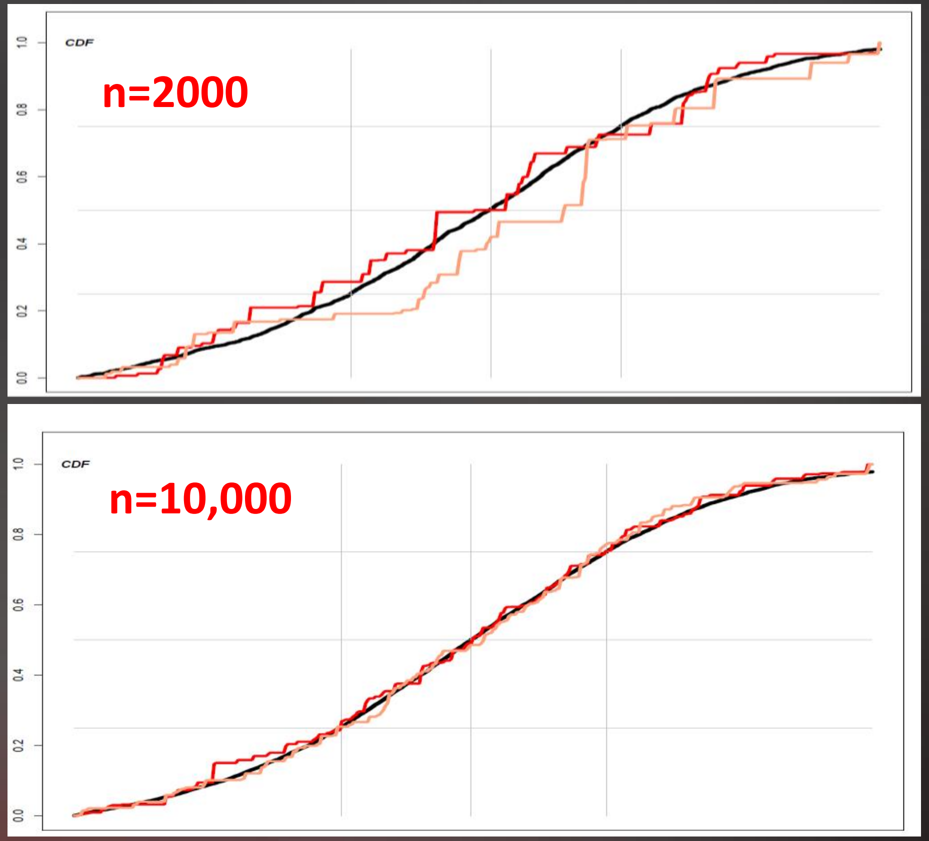


Dataset Size in Diff. Privacy

The intuition for this lies in the way differential privacy operates.

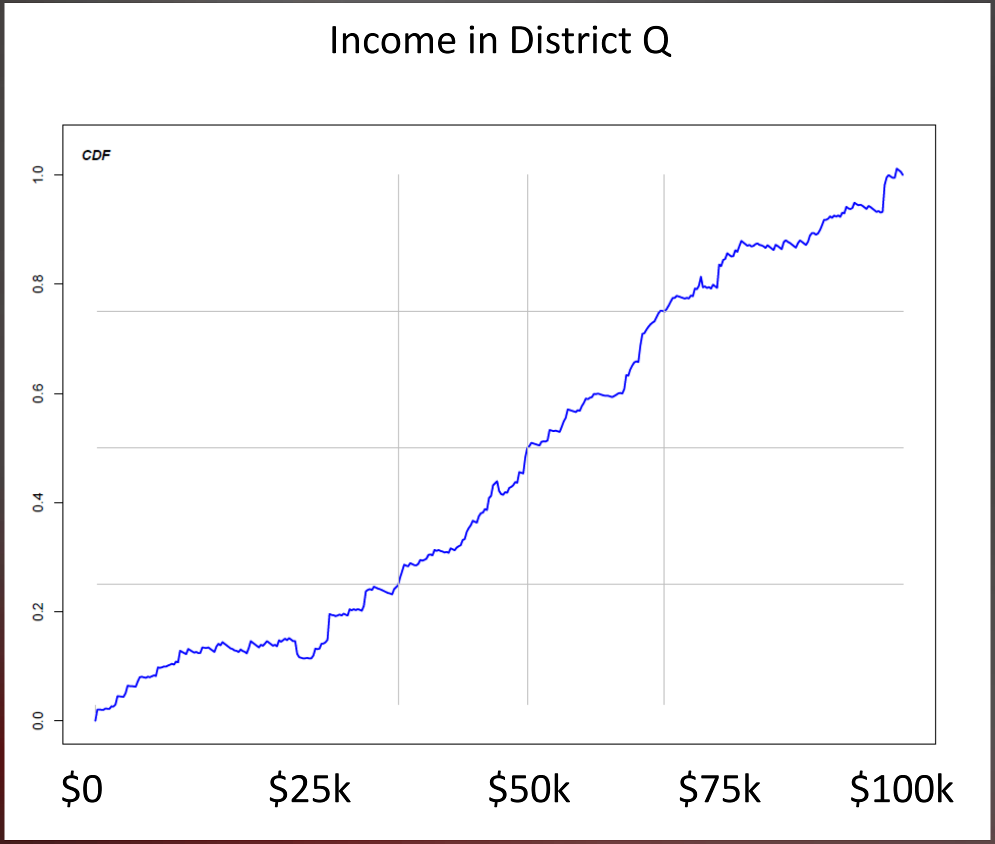
Statistics normally draw on each observation in a dataset only once. Thus, each individual's contribution to the statistic is 1/n. Differential privacy requires that enough random noise be added to obscure the contribution of any individual, so privately releasing this statistic requires noise on the order of 1/n. As n increases, the magnitude of the noise you need to add decreases.

Large ε, 0.300



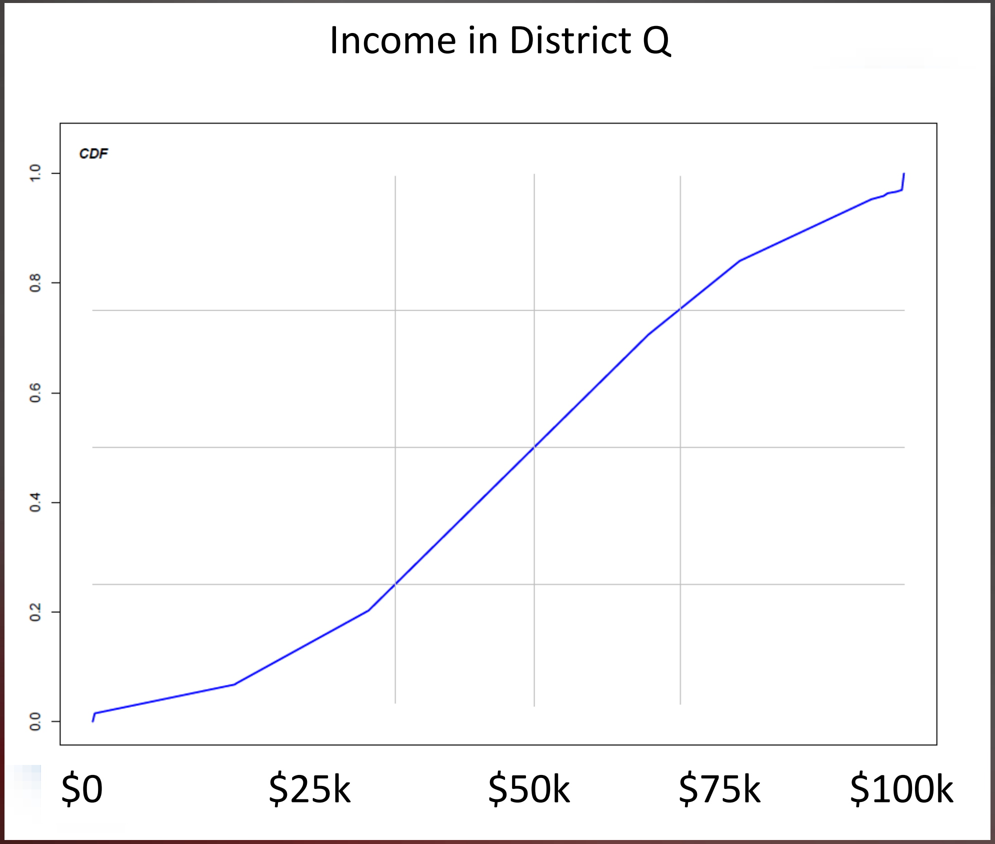
Differential Privacy in CDFs

Now let's return to the ε knob. Gertrude is given 3 DP-CDFs, each with different levels of ε. ADD: tighter and looser DP-CDF



Differential Privacy in CDFs

Now let's return to the ε knob. Gertrude is given 3 DP-CDFs, each with different levels of ε. ADD: tighter and looser DP-CDF



Differential Privacy in CDFs

Now let's return to the ε knob. Gertrude is given 3 DP-CDFs, each with different levels of ε. ADD: tighter and looser DP-CDF