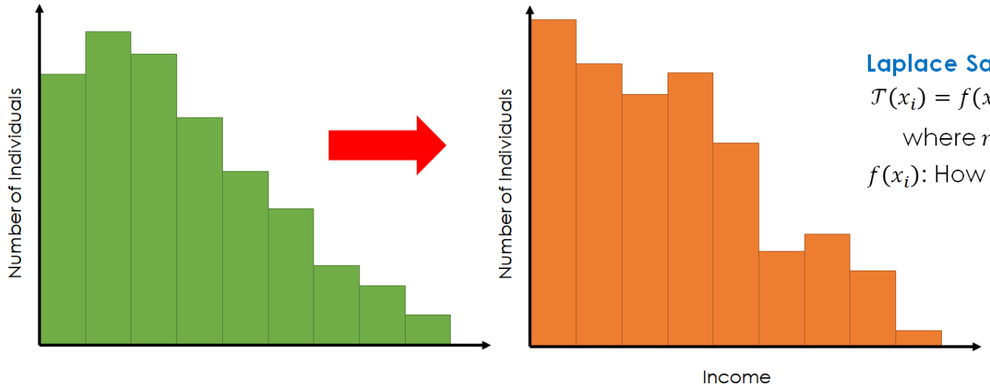
DATA PRIVACY AND ANONYMIZATION IN R

Laplace Sanitizer

Claire McKay Bowen
Postdoctoral Researcher, Los Alamos National Laboratory

Laplace Sanitizer



Laplace Sanitizer

$\mathcal{T}(x_i) = f(x_i) + r_i$

where r_i are drawn i.i.d. from $Lap\left(\frac{S(f)}{\epsilon}\right)$

$f(x_i)$: How many counts?

Male Fertility Data: Prepping Data

```
> fertility %>%
    count(High_Fevers)
# A tibble: 3 x 2
  High_Fevers Count
        <int> <int>
1          -1     9
2           0    63
3           1    28

> # Old: Set Value of Epsilon
> eps <- 0.01 / 2
> # GS of Counts
> gs.count <- 1

> # Set Value of Epsilon
> eps <- 0.01
```

Male Fertility Data: Applying the Laplace mechanism

```
# Apply the Laplace mechanism and set.seed(42)
> set.seed(42)
> fever1 <- rdoublex(1, 9, gs.count / eps) %>%
    max(0)
> fever2 <- rdoublex(1, 63, gs.count / eps) %>%
    max(0)
> fever3 <- rdoublex(1, 28, gs.count / eps) %>%
    max(0)

> fever <- c(fever1, fever2, fever3)

# Normalize noise
> normalized <- (fever/sum(fever)) * (nrow(fertility))
# Round the values
> round(normalized)
[1] 24 76  0
```

Male Fertility Data: Generating Synthetic Data

```
> rep(-1, 24)
 [1] -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1

> rep(0, 76) %>%
    head()
[1] 0 0 0 0 0 0
```

DATA PRIVACY AND ANONYMIZATION IN R

Let's practice!

DATA PRIVACY AND ANONYMIZATION IN R

Differential Privacy (DP) Parametric Approaches

Claire McKay Bowen
Postdoctoral Researcher, Los Alamos National Laboratory

Male Fertility Data

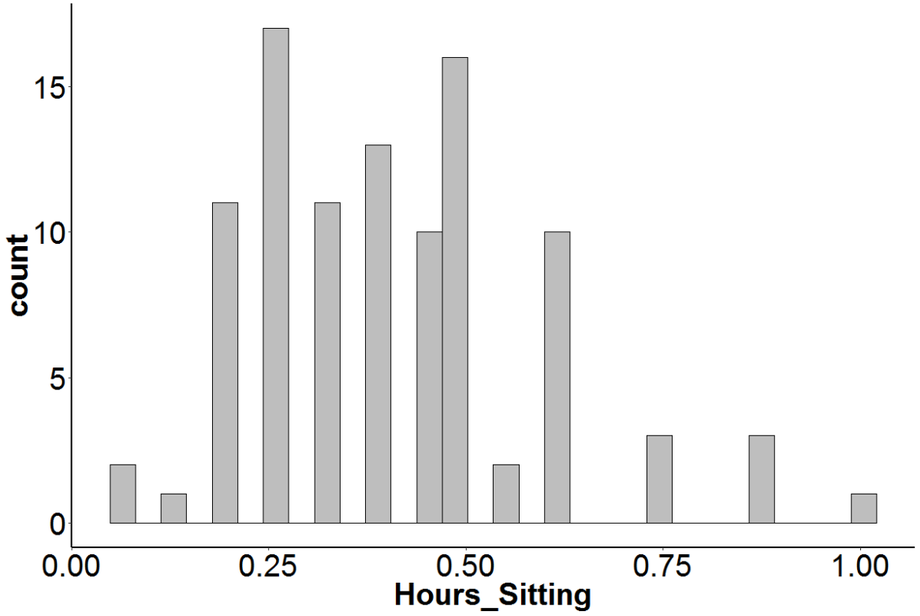
```
> library(dplyr)
> library(smoothmest)

> fertility

# A tibble: 100 x 10
   Season   Age Child_Disease Accident_Trauma Surgical_Intervention
    <dbl> <dbl>         <int>            <int>                <int>
 1  -0.33  0.69             0                1                    1
 2  -0.33  0.94             1                0                    1
 3  -0.33  0.50             1                0                    0
 4  -0.33  0.75             0                1                    1
 5  -0.33  0.67             1                1                    0
 6  -0.33  0.67             1                0                    1
 7  -0.33  0.67             0                0                    0
 8  -0.33  1.00             1                1                    1
 9   1.00  0.64             0                0                    1
10   1.00  0.61             1                0                    0
# ... with 90 more rows, and 5 more variables: High_Fevers <int>,
#   Alcohol_Freq <dbl>, Smoking <int>, Hours_Sitting <dbl>, Diagnosis <int>
```

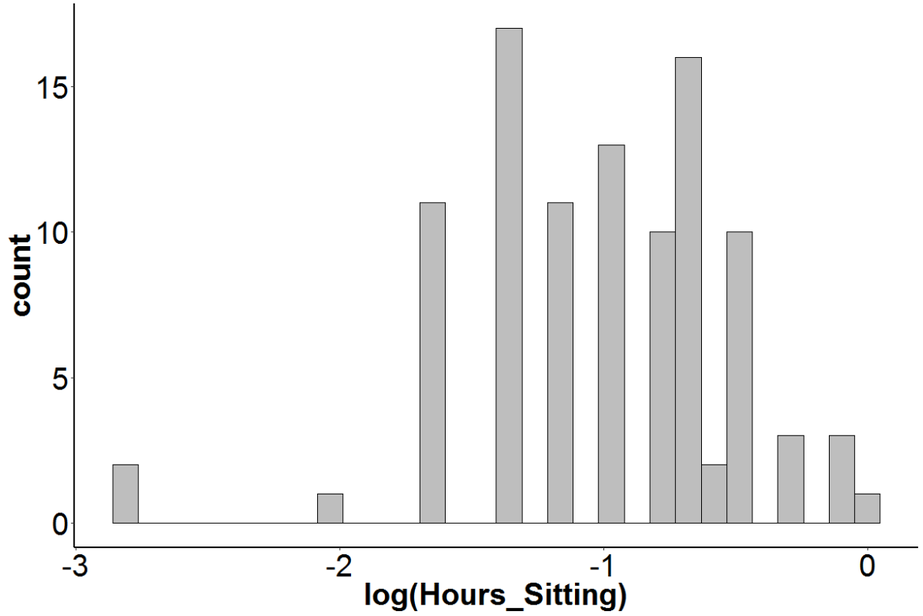
Generating DP Synthetic Data Part 1

Sampling from a Binomial Distribution

```
> fertility %>%
    summarise_at(vars(Child_Disease), mean)

# A tibble: 1 x 1
  Child_Disease
          <dbl>
1          0.87

> set.seed(42)
> rdoublex(1, 0.87, (1 / 100) / 0.1)
[1] 0.8898337

> set.seed(42)
> child.disease <- rbinom(100, 1, 0.89)

> sum(child.disease)
[1] 84
```

Examining the Data





Generating DP Synthetic Data Part 2

Sampling from a Normal Distribution

```
> fertility %>%
    mutate(Hours_Sitting = log(Hours_Sitting)) %>%
    summarise_at(vars(Hours_Sitting), funs(mean, var))

# A tibble: 1 x 2
       mean       var
      <dbl>     <dbl>
1 -1.012244 0.2548017

> set.seed(42)
> rdoublex(1, -1.01, (1 / 100) / 0.01 / 2)
[1] -0.9108316
> rdoublex(1, 0.25, (1 / 100)^2 / 0.01 / 2)
[1] 0.2514175
```

Generating DP Synthetic Data Part 3

Sampling from a Normal Distribution

```
> set.seed(42)
> hours.sit <- rnorm(100, -0.91, sqrt(0.25))

> hours.sit <- exp(hours.sit)

> hours.sit[hours.sit < 0] <- 0
> hours.sit[hours.sit > 1] <- 1

> hours.sit %>%
    head()
[1] 0.3115892 1.0000000 0.6662523 0.4659892 0.3625910 1.0000000
```

DATA PRIVACY AND ANONYMIZATION IN R

Let's practice!

DATA PRIVACY AND ANONYMIZATION IN R

Wrap Up

Claire McKay Bowen
Postdoctoral Researcher, Los Alamos National Laboratory

Chapter 1: Introduction to Data Privacy

- Removing Identifiers
- Generalization
- Top and Bottom coding
- Generating Synthetic Data

Chapter 2: Introduction to Differential Privacy

- Privacy Budget
- Global Sensitivity
- Laplace mechanism

Chapter 3: Differentially Private Properties

- Sequential Composition
- Parallel Composition
- Post-processing
- Impossible and Inconsistent Answers

Chapter 4: Differentially Private Data Synthesis

- Laplace sanitizer
- Parametric approaches

More on Data Privacy

Issues

- Complex solutions for complex data
- Biasing inferences

Other Topics

- Other versions of differential privacy
- Differential privacy methods for specific data types or analyses

DATA PRIVACY AND ANONYMIZATION IN R

Thank you!