DATA PRIVACY AND ANONYMIZATION IN R

Introduction to Anonymization (I)

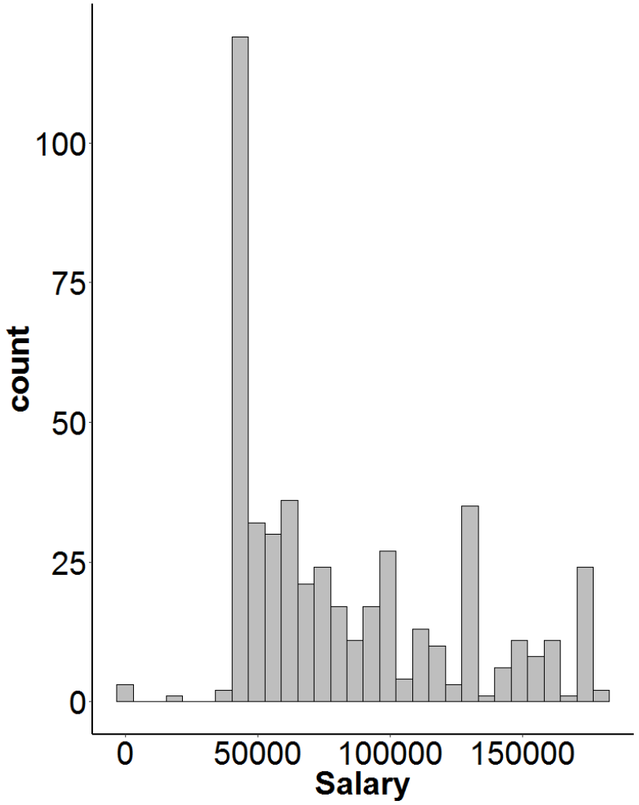
Claire McKay Bowen
Postdoctoral Researcher, Los Alamos National Laboratory

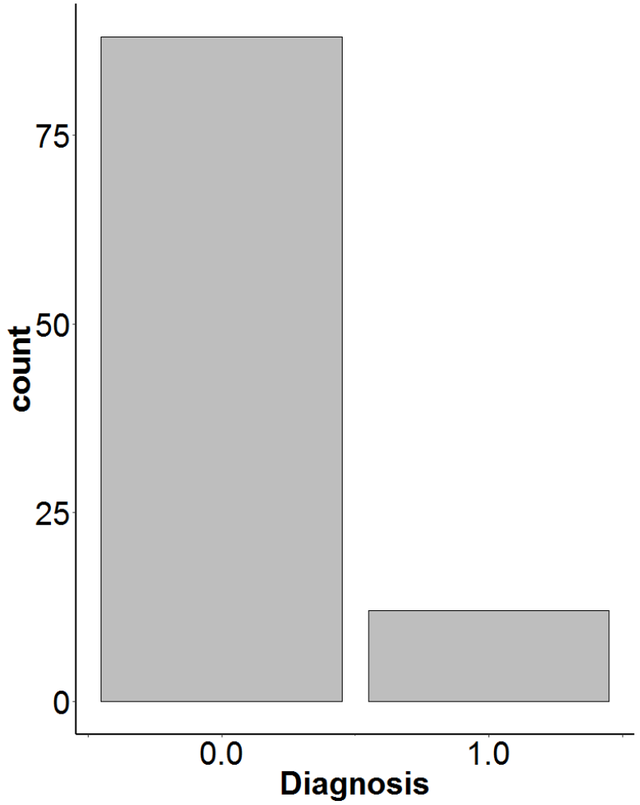

MEDICAL
MEDICAL
MEDICAL
MEDICAL
MEDICAL
MEDICAL
Health Care
Doctor
Hospital
Pharmacist
Nurse
Dentist
First Aid
Surgeon
Emergency

Course Outline

- **Chapter 1:** removing identifiers and generating synthetic data
- **Chapter 2:** differential privacy and Laplace mechanism
- **Chapter 3:** differentially private properties
- **Chapter 4:** differentially private data synthesis

White House Salary and Fertility data sets





The White House Salary data

```
> library(dplyr)

> whitehouse

# A tibble: 469 x 5
                   Name   Status Salary     Basis
                  <chr>    <chr>  <dbl>     <chr>
 1      Abrams, Adam W. Employee  66300 Per Annum
 2        Adams, Ian H. Employee  45000 Per Annum
 3      Agnew, David P. Employee  93840 Per Annum
 4        Albino, James Employee  91800 Per Annum
 5 Aldy, Jr., Joseph E. Employee 130500 Per Annum
 6     Alley, Hilary J. Employee  42000 Per Annum
 7 Amorsingh, Lucius L. Employee  56092 Per Annum
 8  Anderson, Amanda D. Employee  60000 Per Annum
 9 Anderson, Charles D. Employee  51000 Per Annum
10     Andrias, Kate E. Employee 130500 Per Annum
# ... with 459 more rows, and 1 more variables: Title <chr>
```

Removing Identifiers and Rounding

Removing Identifiers

```
> whitehouse %>%
    mutate(Name = 1:469)
```

Rounding

```
> whitehouse %>%
    mutate(Salary = round(Salary, digits = -3))
```

DATA PRIVACY AND ANONYMIZATION IN R

Let's practice!

DATA PRIVACY AND ANONYMIZATION IN R

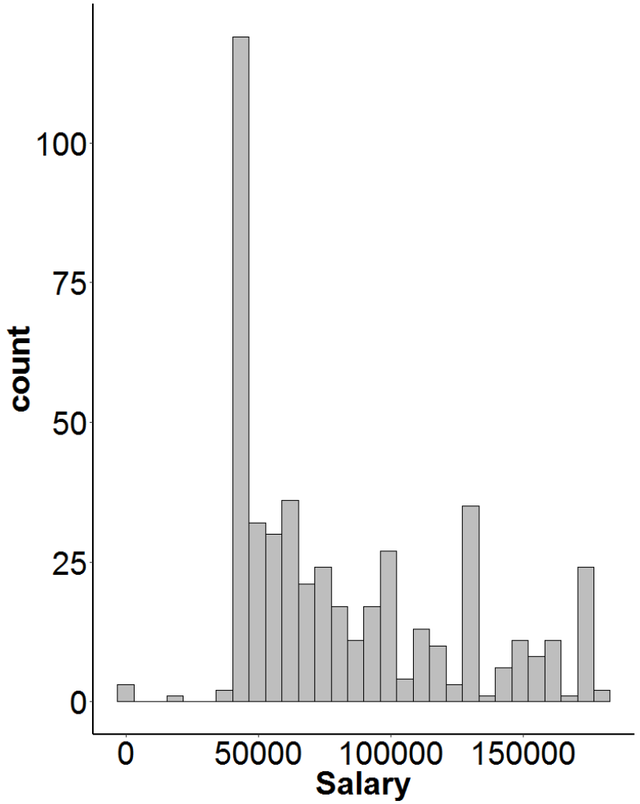
Introduction to Anoynymization (II)

Claire McKay Bowen
Postdoctoral Researcher, Los Alamos National Laboratory

The White House Salary data

```
> whitehouse
# A tibble: 469 x 5
                    Name   Status Salary     Basis
                   <chr>    <chr>  <dbl>     <chr>
 1      Abrams, Adam W. Employee  66300 Per Annum
 2        Adams, Ian H. Employee  45000 Per Annum
 3      Agnew, David P. Employee  93840 Per Annum
 4        Albino, James Employee  91800 Per Annum
 5 Aldy, Jr., Joseph E. Employee 130500 Per Annum
 6     Alley, Hilary J. Employee  42000 Per Annum
 7 Amorsingh, Lucius L. Employee  56092 Per Annum
 8  Anderson, Amanda D. Employee  60000 Per Annum
 9 Anderson, Charles D. Employee  51000 Per Annum
10     Andrias, Kate E. Employee 130500 Per Annum
# ... with 459 more rows, and 1 more variables: Title <chr>
```

Histogram of Salaries



Generalization

```
> whitehouse.gen <- whitehouse %>%
    mutate(Salary = ifelse(Salary < 100000, 0, 1))

> whitehouse.gen
# A tibble: 469 x 5
                   Name   Status Salary     Basis
                  <chr>    <chr>  <dbl>     <chr>
 1      Abrams, Adam W. Employee      0 Per Annum
 2        Adams, Ian H. Employee      0 Per Annum
 3      Agnew, David P. Employee      0 Per Annum
 4        Albino, James Employee      0 Per Annum
 5 Aldy, Jr., Joseph E. Employee      1 Per Annum
 6     Alley, Hilary J. Employee      0 Per Annum
 7 Amorsingh, Lucius L. Employee      0 Per Annum
 8  Anderson, Amanda D. Employee      0 Per Annum
 9 Anderson, Charles D. Employee      0 Per Annum
10     Andrias, Kate E. Employee      1 Per Annum
# ... with 459 more rows, and 1 more variables: Title <chr>
```

Top Coding

```
whitehouse.top <- whitehouse %>%
    mutate(Salary = ifelse(Salary >= 165000, 165000, Salary))

> whitehouse.top %>%
    filter(Salary >= 165000)
# A tibble: 27 x 5
                   Name   Status Salary     Basis
                  <chr>    <chr>  <dbl>     <chr>
 1    Axelrod, David M. Employee 165000 Per Annum
 2    Barnes, Melody C. Employee 165000 Per Annum
 3     Bauer, Robert F. Employee 165000 Per Annum
 4     Brennan, John O. Employee 165000 Per Annum
 5  Brown, Elizabeth M. Employee 165000 Per Annum
 6    Browner, Carol M. Employee 165000 Per Annum
 7    Cutter, Stephanie Employee 165000 Per Annum
 8   Donilon, Thomas E. Employee 165000 Per Annum
 9     Emanuel, Rahm I. Employee 165000 Per Annum
10 Favreau, Jonathan E. Employee 165000 Per Annum
# ... with 17 more rows, and 1 more variables: Title <chr>
```

Quick intro to ...

- `count()`
- `summarise_at()`

count()

```
> whitehouse %>%
+     count(Status)
# A tibble: 3 x 2
  Status                    n
  <chr>                 <int>
1 Detailee                 31
2 Employee                437
3 Employee (part-time)      1
```

count()

```
> whitehouse %>%
+     count(Status, Title, sort = TRUE)
# A tibble: 279 x 3
   Status   Title                                              n
   <chr>    <chr>                                          <int>
 1 Employee STAFF ASSISTANT                                   23
 2 Employee RECORDS MANAGEMENT ANALYST                        15
 3 Employee ANALYST                                           10
 4 Employee SPECIAL ASSISTANT TO THE PRESIDENT AND ASSO…      10
 5 Employee SPECIAL ASSISTANT TO THE PRESIDENT FOR LEGI…      10
 6 Employee ASSOCIATE DIRECTOR                                 9
 7 Employee SENIOR ANALYST                                     8
 8 Employee ASSISTANT DIRECTOR                                 7
 9 Employee SPECIAL ASSISTANT                                  7
10 Employee ASSISTANT SHIFT LEADER                             6
# ... with 269 more rows
```

summarise_at()

```
> whitehouse %>%
    summarise_at(vars(Salary), sum)
# A tibble: 1 x 1
    Salary
     <dbl>
1 38796307
```

summarise_at()

```
> whitehouse %>%
    summarise_at(vars(Salary), funs(mean, sd))
# A tibble: 1 x 2
      mean       sd
     <dbl>    <dbl>
1 82721.34 41589.43
```

DATA PRIVACY AND ANONYMIZATION IN R

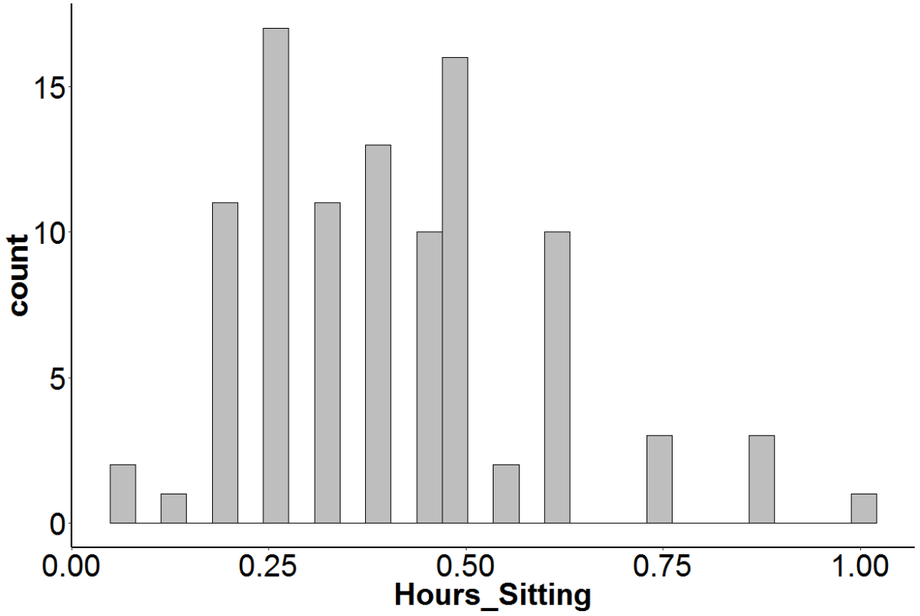
Let's practice!

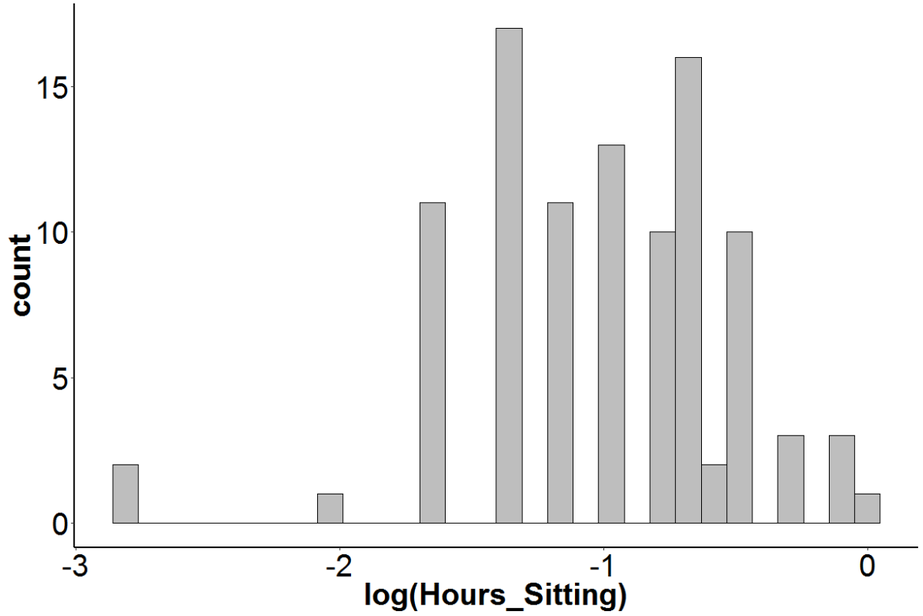
DATA PRIVACY AND ANONYMIZATION IN R

Data Synthesis

Claire McKay Bowen
Postdoctoral Researcher, Los Alamos National Laboratory

Probability Distributions





Male Fertility Data

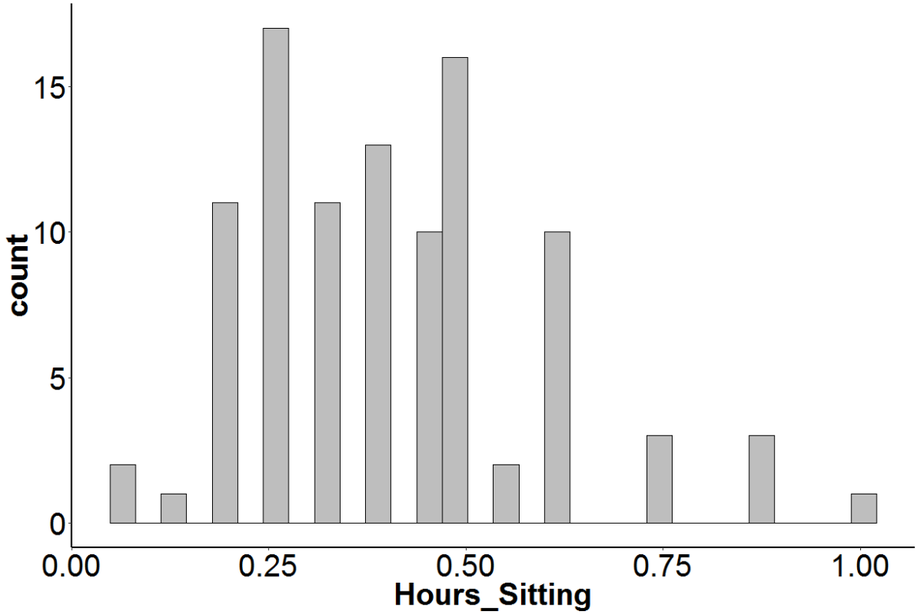
```
> library(dplyr)

> 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>
```

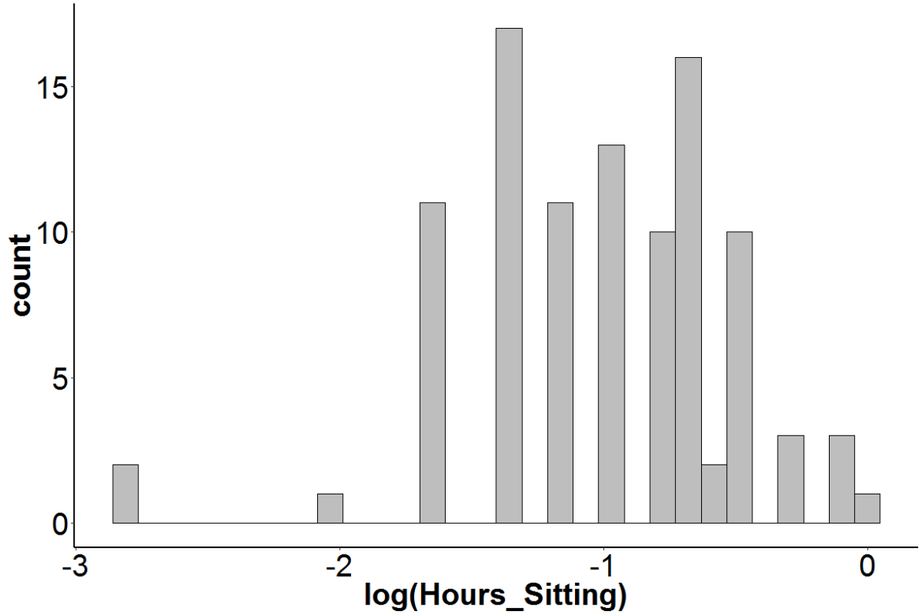
Generating Synthetic Data Part 1

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)
> child.disease <- rbinom(100, 1, 0.87)

> sum(child.disease)
[1] 83
```

Examining the Data





Generating Synthetic Data Part 2

Sampling from a Normal Distribution

```
> fert <- fertility %>%
    mutate(Hours_Sitting = log(Hours_Sitting))

> fert %>%
    summarise_at(vars(Hours_Sitting), funs(mean, sd))

# A tibble: 1 x 2
    mean        sd
   <dbl>     <dbl>
1 -1.012244 0.5047788

> set.seed(42)
> hours.sit <- rnorm(100, -1.01, 0.50)

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

How to Handle Improper Values

Hard Bounding

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

> range(hours.sit)
[1] 0.0815495 1.0000000
```

DATA PRIVACY AND ANONYMIZATION IN R

Let's practice!