MAT 4376 Foundations of Data Privacy

Rafał Kulik

Department of Mathematics and Statistics (University of Ottawa)

Lectures 7, 8, 9

27, 29 January and 3 February 2025

1. Measuring data utility and information loss
   - Introduction
   - Some general tools
   - General utility measures for categorical variables
   - General utility measures for continuous variables

During these lectures we will learn:

- data utility metrics for categorical variables;
- data utility metrics for continuous variables.

Some resources

- A blog on SDC: `https://sdcpractice.readthedocs.io/en/latest/measure_risk.html`.
- MSc thesis of Chang Qu.
- PhD thesis of Devyani Biswal.

Introduction

SDC is a **trade-off** between risk of disclosure and loss of data utility and seeks to minimize the latter, while reducing the risk of disclosure to an acceptable level. Data utility in this context means the usefulness of the anonymized data for statistical analyses by end users as well as the validity of these analyses when performed on the anonymized data.
In order to make a trade-off between minimizing disclosure risk and maximizing utility of data for end users, it is necessary **to measure the utility of the data after anonymization and compare it with the utility of the original data**.
Terminology: **Information loss** is the inverse of data utility: the larger the data utility after anonymization, the smaller the information loss.

Introduction

If the microdata to be anonymized is based on a sample, the data will incur a **sampling error**.
Other errors may be present in the data, such as nonresponse error.
The methods discussed here only measure the information loss caused by the anonymization process relative to the original sample data and do not attempt to measure the error caused by other sources. Some of these data utility measures allow for statistical inference.
The main set up: we have the original database $X = (X_1, \ldots, X_n)$ and an anonymized database $Y = (Y_1, \ldots, Y_m)$. Note that in general $n \neq m$.
The observations X_j and Y_j may be univariate or multivariate,
$X_j = (X_{j1}, \ldots, X_{jp})$, $Y_j = (Y_{j1}, \ldots, Y_{jp})$.

Introduction

- Some of the data utility measures are used as follows: calculate the utility measure for the original data set, then do the same for the anonymized data set. A *significant* change indicates data utility loss.
- Similarly, we can calculate the data utility measures for one anonymized dataset and compare to another anonymized data set, to assess which anonymization yields better data utility.
- Some of the data utility measures compare the original and the anonymized data set.

Mean Squared Error of an estimator

Let $X_1, \ldots, X_n$ be a random sample and $\widehat{\theta} = T(X_1, \ldots, X_n)$ be an estimator of the parameter θ of interest. For example, $T(x_1, \ldots, x_n) = \bar{x} := (x_1 + \cdots + x_n)/n$ yields sample mean.

- Define $\mathrm{bias}(\widehat{\theta}) = \mathbb{E}[\widehat{\theta}] - \theta$. If the bias is zero, we call $\widehat{\theta}$ an *unbiased estimator*.
- Mean Squared Error of the estimator $\widehat{\theta}$ is defined as

$$\mathrm{MSE}(\widehat{\theta}) = \mathbb{E}\left[(\widehat{\theta} - \theta)^2\right] .$$

It turns out that (assignment)

$$\mathrm{MSE}(\widehat{\theta}) = \mathrm{Var}(\widehat{\theta}) + (\mathrm{bias}(\widehat{\theta}))^2 .$$

Idea: Calculate MSE for the original and the anonymized data. The bigger the difference, the less data utility.

Kolmogorov-Smirnov statistics

Let us start with a sample $X_1, \ldots, X_n$ coming from the population with mean μ and variance σ^2. Let $\bar{X}$ be the sample mean. Then

$$\mathbb{E}[\bar{X}] = \mu \,, \quad \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n} \,.$$

Hence, we have the Central Limit Theorem (CLT)

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1) \,.$$

We can use this information to construct the confidence interval for mean:

$$\mathbb{P}\left(\mu \in L(\mathbf{X}), U(\mathbf{X})\right) = 1 - \alpha \,,$$

where $L(\mathbf{X}) = \bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ and $L(\mathbf{X}) = \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, with $z_{\alpha/2}$ being the normal percentile. We can also use this information for hypothesis testing.

Kolmogorov-Smirnov statistics

Now, assume that instead of estimating the mean, we want to estimate $F(x_0)$, the cumulative distribution function at point x_0. Define $Y_j = \mathbb{1}\{X_j \leq x_0\}$. Note that

$$\mathbb{E}[Y_j] = F(x_0)\,, \quad \mathrm{Var}(Y_j) = F(x_0)(1 - F(x_0))\,.$$

Define the *empirical (cumulative) distribution function*

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}\{X_j \leq x\}\,, \quad x \in \mathbb{R}\,.$$

Then

$$\mathbb{E}[\widehat{F}_n(x_0)] = F(x_0)\,, \quad \mathrm{Var}(\widehat{F}_n(x_0)) = \frac{F(x_0)(1 - F(x_0))}{n}\,.$$

Kolmogorov-Smirnov statistics

We can automatically conclude CLT:

$$\sqrt{n}\left\{\widehat{F}_n(x_0) - F(x_0)\right\} \xrightarrow{\mathrm{d}} \mathcal{N}(0, F(x_0)(1 - F(x_0)) .$$

From this we could construct confidence interval, but we have a problem - it will involve $F(x_0)$ that we are looking for! We do a usual trick, by estimating $F(x_0)$. The confidence interval is

$$\widehat{F}_n(x_0) \pm z_{\alpha/2}\sqrt{\frac{\widehat{F}_n(x_0)(1 - \widehat{F}_n(x_0))}{n}} .$$

We can still do hypothesis testing $H_0 : F(x_0) = p$, where $p \in (0, 1)$.

Kolmogorov-Smirnov statistics

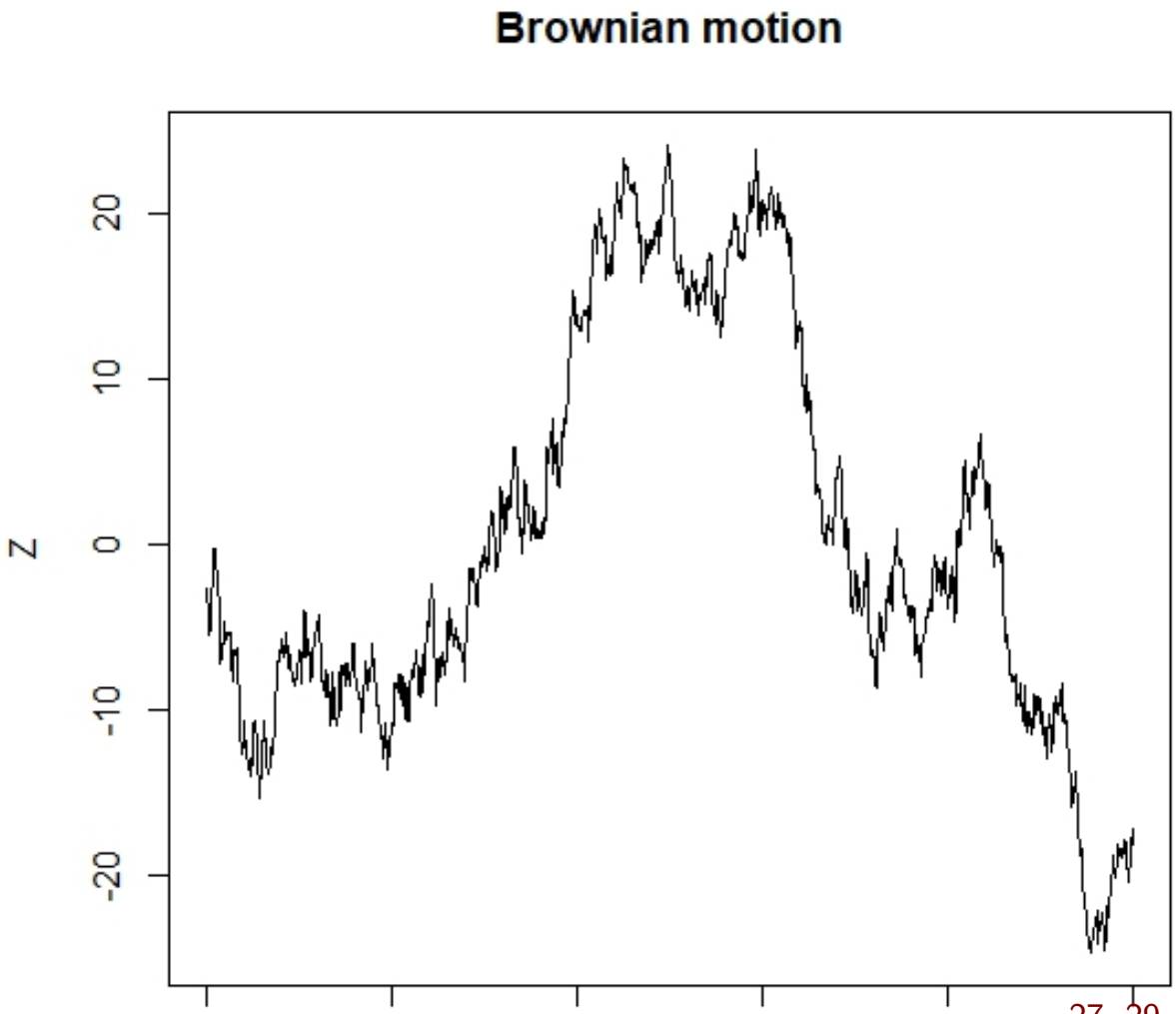
However, what we are really interested in is estimation of the entire CDF $F(x)$. Then, we need to treat $\widehat{F}_n(x)$ as a function of x. This leads to *Functional Central Limit Theorem*:

$$\sqrt{n}\left\{\widehat{F}_n(x) - F(x)\right\} \xrightarrow{\mathrm{d}} W(F(x)),$$

where W is a *Brownian motion*.

Kolmogorov-Smirnov statistics

```
X=rnorm(1000); Z=cumsum(X);
```



Kolmogorov-Smirnov statistics

Define the Kolmogorov-Smirnov statistics:

$$D_n = \sup_{x \in \mathbb{R}} |\widehat{F}_n(x) - F(x)| \ .$$

Then

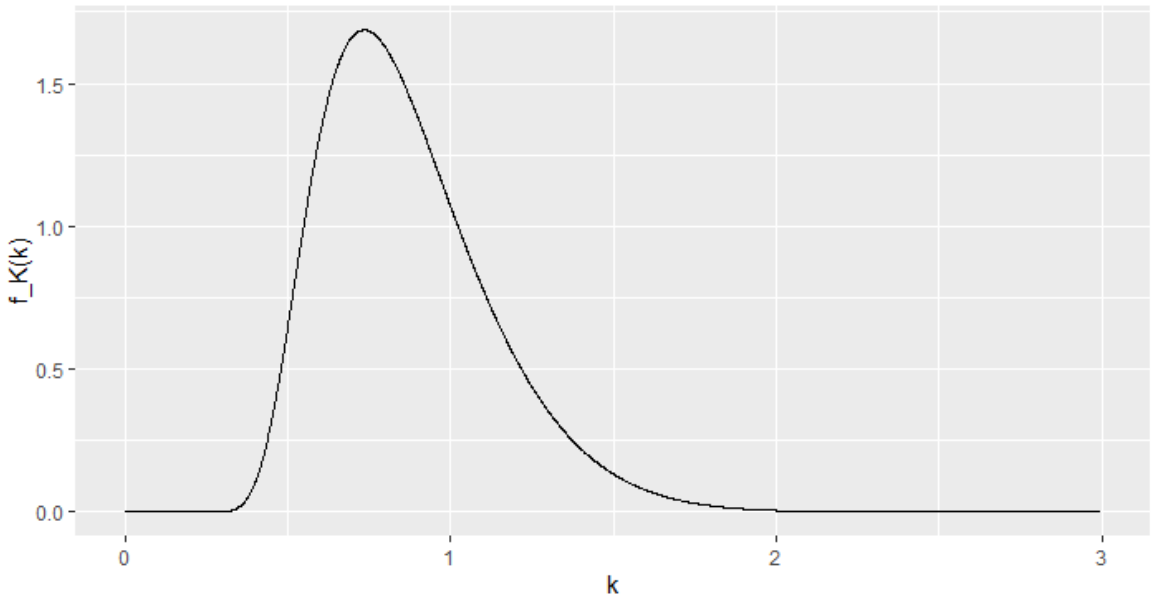
$$\sqrt{n} D_n \xrightarrow{\mathrm{d}} \sup_{x \in \mathbb{R}} W(F(x)) = \sup_{t \in [0,1]} W(t) =: K \ .$$

The distribution of a random variable on the right hand side is non-trivial. Indeed,

$$\mathbb{P}(K \leq x) = 1 - 2 \sum_{k=1}^{\infty} (-1)^{k-1} e^{-2k^2 x^2} \ .$$

Kolmogorov-Smirnov test

Assume now we want to test if our data come from the normal distribution with mean μ and variance σ^2. Then we calculate D_n with $F = \Phi_{\mu,\sigma^2}$, the CDF of the $\mathcal{N}(\mu, \sigma^2)$. We reject the null hypothesis whenever $\sqrt{n}D_n > q_{1-\alpha}$, where $q_{1-\alpha}$ is the quantile of K.



Kolmogorov-Smirnov test

We could use the KS test as follows: test for a particular distribution based on the original and anonymized data. If one null hypothesis is rejected and another one is not, then it indicates loss of data utility.

Kolmogorov-Smirnov test - two samples

Assume now that we have data $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ coming from CDFs F and G, respectively. We want to test whether they come from the same distribution. Define

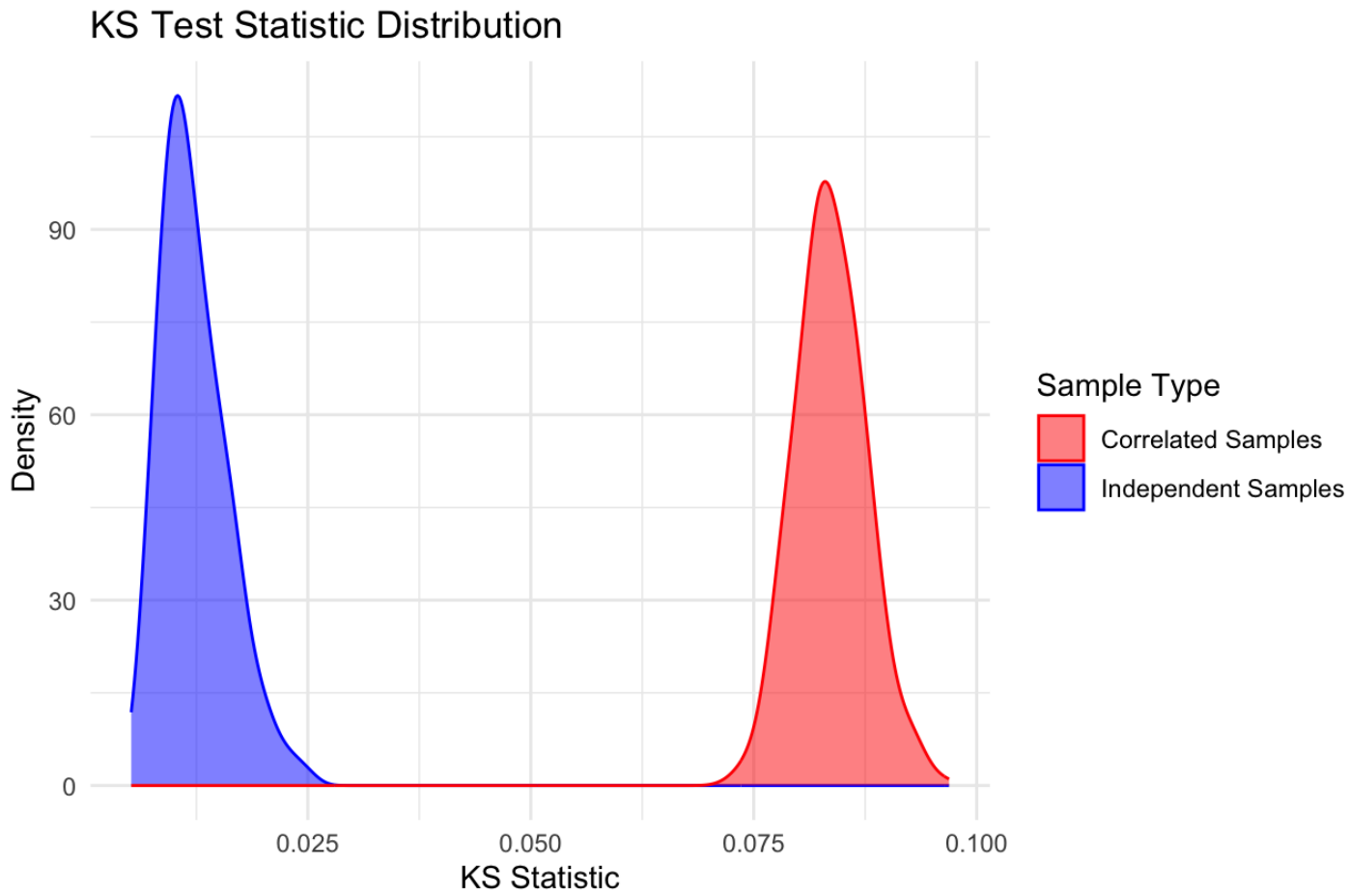
$$D_{n,m} = \sup_{x \in \mathbb{R}} |\widehat{F}_n(x) - \widehat{G}_m(x)| \,,$$

where $\widehat{G}_m$ is the empirical CDF based on $Y_1, \ldots, Y_m$. Consider the null hypothesis $H_0 : F = G$ at the level $1 - \alpha$. Then we reject null whenever

$$D_{n,m} \geq \sqrt{-\ln(\alpha/2)\frac{1 + m/n}{2m}} \,.$$

Kolmogorov-Smirnov test - two samples - be careful!!!

Important: this test is valid if both samples are independent. In the data privacy context these the two samples are usually dependent and the test **not valid**. Yet, it is implemented in many data privacy software ...



Distance between probability distribution

Let P and Q be two probability distributions. For example, if X and Y are random variables, then P and Q could represent $P(A) = \mathbb{P}(X \in A)$ and $Q(A) = \mathbb{P}(Y \in A)$. If $A = (-\infty, x]$, $P(A)$ becomes $F(x)$, the CDF.
We want to measure the distance between P and Q. One possibility is the Kolmogorov-Smirnov distance:

$$D_{KS}(P\|Q) := \sup_A |P(A) - Q(A)| .$$

In one dimensional case we can reduce it to

$$\sup_{x\in\mathbb{R}} |F(x) - G(x)| .$$

The statistical procedures for the latter were presented above.

Kullback–Leibler divergence

Define

$$D_{KL}(P\|Q) := \sum_x f(x) \log\left(\frac{f(x)}{g(x)}\right)$$

or

$$D_{KL}(P\|Q) := \int f(x) \log\left(\frac{f(x)}{g(x)}\right) dx \ ,$$

where f and g are probability mass functions (discrete case) or density functions (continuous case) of random variables X and Y, respectively. The above quantity is called the Kullback–Leibler (KL) divergence, or relative entropy, or ℓ-divergence.

Kullback–Leibler divergence

Note that

$$D_{KL}(P\|Q) = \mathbb{E}\left[\frac{f(X)}{g(X)}\right] .$$

Obviously, if $P = Q$, then $D_{KL} = 0$.
Also, D_{KL} is not a distance, since it is not symmetric and does not satisfy the triangle inequality. Furthermore, no upper bound is given.

Trivial measure

The first trivial measure is **the number of records changed per variable**. That is, if X_i, $i = 1, \ldots, n$ are original univariate data and Y_i, $i = 1, \ldots, n$ are the anonymized data, we calculate

$$\sum_{i=1}^{n} \mathbb{1}\{X_j \neq Y_j\} \, .$$

If the dataset is p-dimensional, we evaluate the above expression for all p variables separately. The bigger the value, the less data utility.

Application: Compare the anonymized and the original dataset directly.

Contingency Table

A contingency table is a type of table that displays the frequency distribution of multivariate observations. Each cell in the table represents the count of observations for a specific combination of categories. in case of continuous variables one groups them into categories. In what follows,

- Let C be the set of all possible combinations of categorical and grouped continuous variables.
- For each $c \in C$:
  - $td(c)$ is the count of data points in c from the original data. That is

$$td(c) = \sum_{j=1}^{n} \mathbf{1}((X_{j1}, \ldots, X_{jp}) = c).$$

  - $ta(c)$ is the count of data points in c from the anonymized data. That is

$$ta(c) = \sum_{j=1}^{m} \mathbf{1}((Y_{j1}, \ldots, Y_{jp}) = c).$$

Contingency Table

Several standard data utility measures based on the contingency table. This method compares the frequency between groups in both observed and anonymized data.

- **Variance-Weighted Measure (VW)**. It calculates the sum of squared differences between observed and expected counts, weighted by the expected counts:

$$\text{VW} = \sum_{c \in C} \left(\frac{(td(c) - ta(c))^2}{ta(c)} \right) .$$

Contingency Table

- **Jensen-Shannon Divergence (JSD)**. It is a symmetrized version of the Kullback-Leibler divergence. A JSD of 0 means that the two distributions are identical, while a JSD of 1 indicates maximum divergence:

$$\frac{1}{2}\sum_{c\in C} p(c)\log\left(\frac{2p(c)}{p(c)+q(c)}\right)+\frac{1}{2}\sum_{c\in C} q(c)\log\left(\frac{2q(c)}{p(c)+q(c)}\right),$$

where $p(c)=\frac{td(c)}{\sum_{c\in C} td(c)}$ and $q(c)=\frac{ta(c)}{\sum_{c\in C} ta(c)}$.

Contingency Table

- **Kolmogorov-Smirnov Statistic**:

$$\max_{c \in C} \left| \frac{\sum_{i \in c} td(i)}{\sum_{i \in C} td(i)} - \frac{\sum_{i \in c} ta(i)}{\sum_{i \in C} ta(i)} \right| .$$

  The closer the value to zero indicates better utility.

Some packages provide statistical goodness of fit tests for Jensen-Shannon Divergence and Kolmogorov-Smirnov Statistic. They seem to be incorrect, since the aforementioned statistics are based on dependent data.

Application: Calculate it for the original and the anonymized data. Significant change indicated information loss.

Contingency Table - Example

ID	**Age**	**Income**	**Gender**
1	22	30000	M
2	25	35000	F
3	28	40000	M
4	30	45000	F
5	35	50000	M
6	40	60000	F
7	45	70000	M
8	50	80000	F
9	55	85000	M
10	60	90000	F

Table: Demographic Data of Individuals

Contingency Table - Example

ID	Age	Income	Gender	Group (c)
1	40	60257.33	F	2
2	22	33786.17	M	1
3	30	45652.61	M	1
4	30	40345.76	M	1
5	50	75937.89	F	2
6	60	95796.99	F	2
7	30	49990.02	M	1
8	40	60031.50	F	2
9	28	47308.36	M	1
10	35	49256.03	M	2

Table: Dataset with Grouping: Group 1 - income less than 60,000

Contingency Table - Example

Group (c)	**Anonymized Count ($ta(c)$)**	**Observed Count ($td(c)$)**
1	5	6
2	5	4

Table: Observed Counts for Groups

Contingency Table - Example

Variance-Weighted Measure (VW):

$$VW = \frac{(6-5)^2}{5} + \frac{(4-5)^2}{5} = 0.4.$$

Jensen-Shannon Divergence (JSD): Using $p(c) = \frac{td(c)}{10}$ and $q(c) = \frac{ta(c)}{10}$,

$$\frac{1}{2}\left[\frac{6}{10}\log\left(\frac{12}{11}\right) + \frac{5}{10}\log\left(\frac{10}{11}\right)\right] + \frac{1}{2}\left[\frac{4}{10}\log\left(\frac{8}{9}\right) + \frac{5}{10}\log\left(\frac{10}{9}\right)\right].$$

(Kolmogorov-Smirnov Statistic):

$$\max\left(\left|\frac{6}{10} - \frac{5}{10}\right|, \left|\frac{10}{10} - \frac{10}{10}\right|\right) = 0.1.$$

A measure closer to zero indicates better utility; however, these numbers lack precise interpretations as they are just measures of distance. No proper test statistics are provided.

Contingency Table - some issues

We will illustrate how the introduced distance measures fluctuate based on the method used to group data. Specifically, when a single group is selected, the distance measure will always be zero, regardless of how the data are generated, indicating perfect data utility. When multiple groups are used, even a single difference will lead to a large distance, thus underestimating the utility of the data. As such, **these measures may not be particularly useful when calculate on one anonymized dataset**. They are useful when calculated for, say, two anonymized dataset and compared to indicate which method of anonymization gives better data utility.

Contingency Table - Example II

Given the same dataset as before, we now categorize the income data into three different groups to see how this affects the calculation of our statistical measures (VW, JSD, and KS).

ID	Income	Group by $60,000	New Grouping
1	60257.33	2	2
2	33786.17	1	1
3	45652.61	1	2
4	40345.76	1	2
5	75937.89	2	2
6	95796.99	2	3
7	49990.02	1	2
8	60031.50	2	2
9	47308.36	1	2
10	49256.03	2	2

Table: Income Grouping

Contingency Table - Example II

- Variance-Weighted Measure (VW):

$$VW = \frac{(4-8)^2}{8} + \frac{(10-10)^2}{10} + \frac{(4-2)^2}{2} = 2 + 0 + 4 = 6.$$

- Kolmogorov-Smirnov Statistic:

$$\max\left(\left|\frac{4}{18} - \frac{8}{18}\right|, \left|\frac{10}{18} - \frac{10}{18}\right|, \left|\frac{4}{18} - \frac{2}{18}\right|\right) = \max\left(\frac{4}{18}, 0, \frac{2}{18}\right) = 0.22.$$

Basic statistics

The statistics characterizing the dataset should not change too much after the anonymization. Examples of such statistics are the mean, variance, and covariance and correlation structure of the most important variables in the dataset.

Application: Calculate it for the original and the anonymized data. Significant change indicates information loss in relation to the sample particular statistics.

Basic statistics

Example 1

Our dataset is onedimensional, of size $n = 100$. Calculate confidence interval for the mean. Consider the anonumized dataset. If the anonymized mean belongs to the confidence interval, there is no information loss in relation to the sample mean.

```
X=rnorm(100,50,5) # original data set
# confidence interval
q=qnorm(0.975);
c(mean(X)-q*sd(X)/10,mean(X)+q*sd(X)/10);
# anonymized data
Y=X+rnorm(100,0,1)
mean(Y)
```

Alternatively, test for equality of means. Used **paired test**.

Basic statistics

Let X_1 and X_2 be two random variables with covariance $\mathrm{Cov}(X_1, X_2)$. If Z_1 and Z_2 are independent random variables, independent from X_1 and X_2, then

$$\mathrm{Cov}(X_1 + Z_1, X_2 + Z_2) = \mathrm{Cov}(X_1, X_2) \, .$$

Hence, independent noise addition preserve the covariance between two variables. Correlation is not preserved!

Assume that the original data set has n entries of dimension $p = 2$: $X_j = (X_{j1}, X_{j2})$, $j = 1, \ldots, n$. Then, the the sample covariance coefficient is defined as

$$\frac{1}{n} \sum_{j=1}^{n} (X_{j1} - \bar{X}_1)(X_{j2} - \bar{X}_2) \, ,$$

where $\bar{X}_1$ and $\bar{X}_2$ are sample means for the first and the second variable.

Basic statistics

In what follows we will compare sample covariances and sample correlations for the original and anonymized data.

```
set.seed(105)
X1=rnorm(100,50,5); X2=0.9*X1+rnorm(100,0,1);
cov(X1,X2); cor(X1,X2);
# Anonymized data
Y1=X1+rnorm(100,0,1); Y2=X2+rnorm(100,0,2);
cov(Y1,Y2); cor(Y1,Y2);
```

The results are: 16.3777; 0.97 (original data); 16.47706; 0.87 (anonymized data)

Basic statistics

Some anonymization methods may destroy covariance structure:

```
# Anonymized data
U1=X1+rnorm(100,0,1); U2=sample(X2,100);
cov(U1,U2); cor(U1,U2);
```

The results: 3.76, 0.21.

MSE and coefficient of determination

As indicated before, there are many *Mean Squared Errors*. Here,

$$\text{MSE}_i := \frac{1}{n}\sum_{j=1}^{n}(X_{ji} - Y_{ji})^2, \quad i = 1, \ldots, p\ .$$

Consider also the coefficient of determination given by

$$R_i^2 = 1 - \frac{\sum_{j=1}^{n}(X_{ji} - Y_{ji})^2}{\sum_{j=1}^{n}(X_{ji} - \bar{X}_i)^2}, \quad i = 1, \ldots, p,$$

where $\bar{X}_i$ is the mean of the observed values for the ith variable.

Variance, bias, MSE

Assume that we are interested in estimating a population parameter θ. Let $\widehat{\theta}_X := T(X_1, \ldots, X_n)$ and $\widehat{\theta}_Y := T(Y_1, \ldots, Y_n)$ be estimators of θ based on the original and the anonymized dataset, respectively. Typically, we would have

$$\mathbb{E}[\widehat{\theta}_X] = \mathbb{E}[\widehat{\theta}_Y],$$

but

$$\mathrm{Var}(\widehat{\theta}_X) \leq \mathrm{Var}(\widehat{\theta}_Y) \, .$$

Hence, the estimator based on the anonymized data will have a larger MSE.

Variance, bias, MSE

Example 2

Assume that $X_1, \ldots, X_n$ come from a population with mean μ and variance σ^2. Let $Y_j = X_j + Z_j$, where Z_j are zero-mean, with the variance σ_Z^2, independent from X_j. Then

$$\mathbb{E}[\bar{X}] = \mu = \mathbb{E}[\bar{Y}]$$

and

$$\mathrm{Var}[\bar{X}] = \frac{\sigma^2}{n} \neq \mathrm{Var}[\bar{Y}] = \frac{\sigma^2 + \sigma_Z^2}{n} \ .$$

Variance, bias, MSE

It is possible that noise addition introduces bias. Consider two random variables X and $Y = X + Z$, where Z is independent of X. Assume CDFs of X and Z are continuous and strictly increasing. Then

$$m_X := \text{median}(X) = Q(1/2) \,,$$

where Q is the quantile function. We have

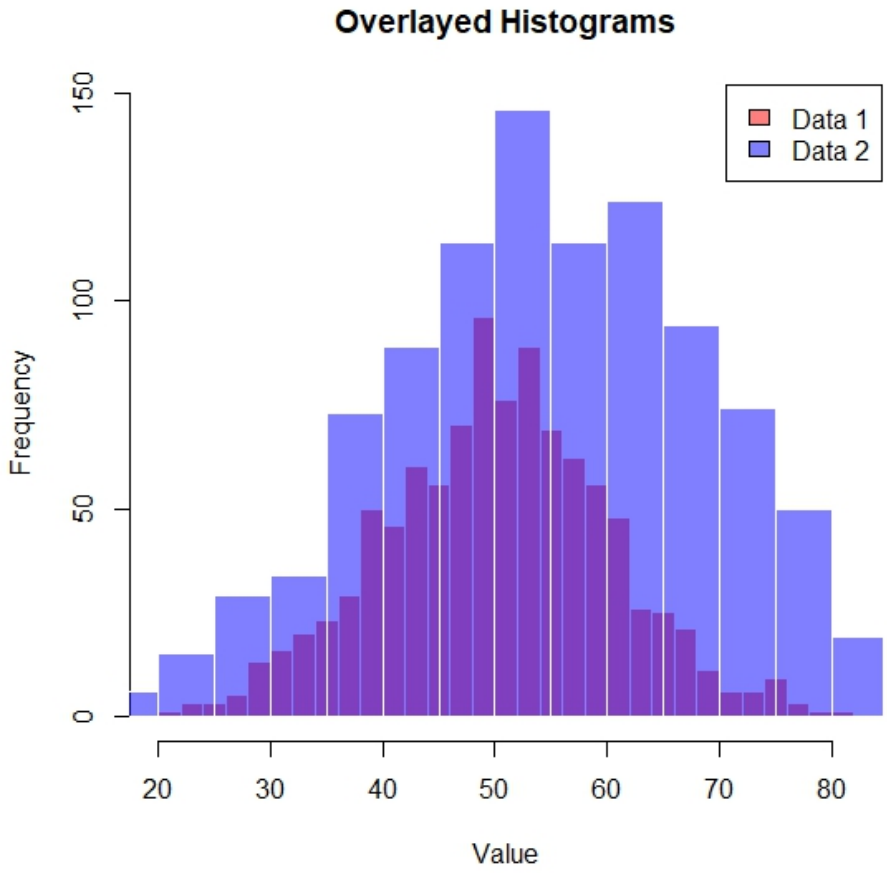
$$\mathbb{P}(X \leq m_X) = 1/2 = \mathbb{P}(X > m_X) \,.$$

In general, it is not true that $m_{X+Z} = m_X + m_Z$ (assignment). Likewise,

$$\text{median}(X_1+Z_1, \ldots, X_n+Z_n) \neq \text{median}(X_1, \ldots, X_n)+\text{median}(Z_1, \ldots, Z_n) \,.$$

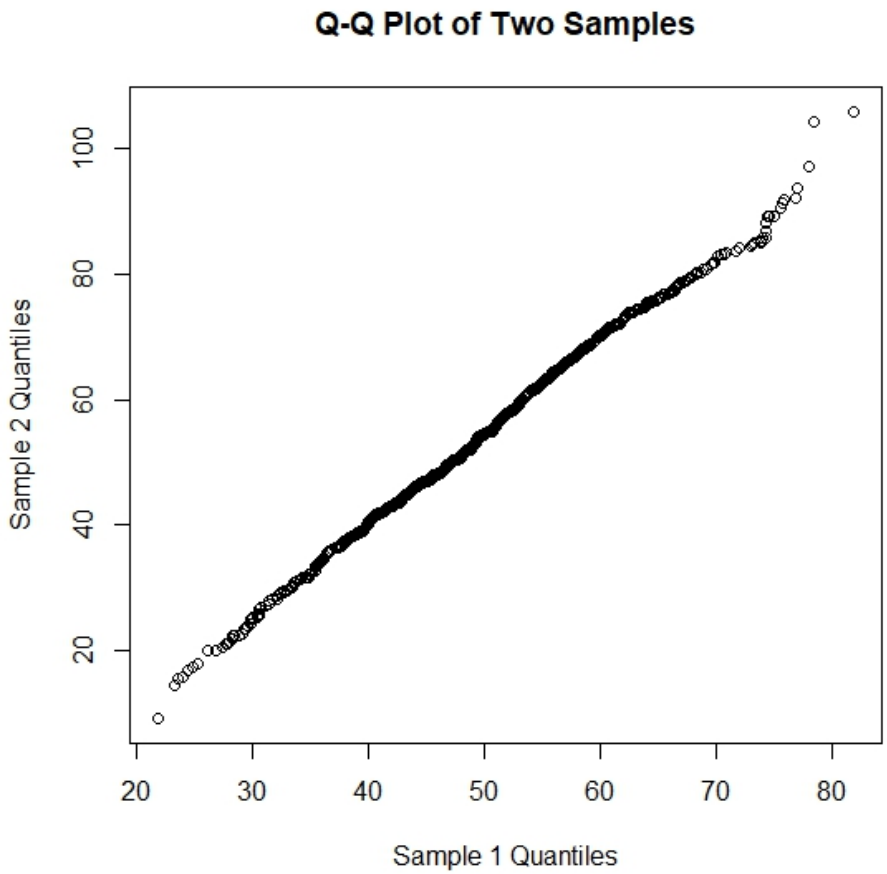
Graphical tools - histogram

We can compare histograms of the original and anonymized data sets. We can also draw a two sample qq-plot.



Graphical tools - QQ-plot

QQ-plot displays pairs $(X_{(j)}, Y_{(j)})$, $j = 1, \ldots, n$, where $X_{(1)} \leq \cdots \leq X_{(n)}$.



Graphical tools - Contour plots

An alternative measure of utility could be based on the comparison of contour plots for dependence structures. Contour plots are a graphical representation used to illustrate the joint distribution of two continuous random variables X and Y. Mathematically, the contour plot is constructed by plotting the level curves of the joint PDF:

$$f_{X,Y}(x, y) = c,$$

where c is a constant representing a specific probability density value. Each contour line corresponds to a different value of c.

Graphical tools - Contour plots

To calculate contour plots for both original and generated data, we use Kernel Density Estimation (KDE) to smooth them. KDE is a nonparametric method for estimating the probability density function of a random variable. For a pair of continuous random variables X and Y, the KDE of the joint PDF is given by:

$$\widehat{f}_{X,Y}(x,y) = \frac{1}{nh_X h_Y} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_X}\right) K\left(\frac{y - Y_i}{h_Y}\right),$$

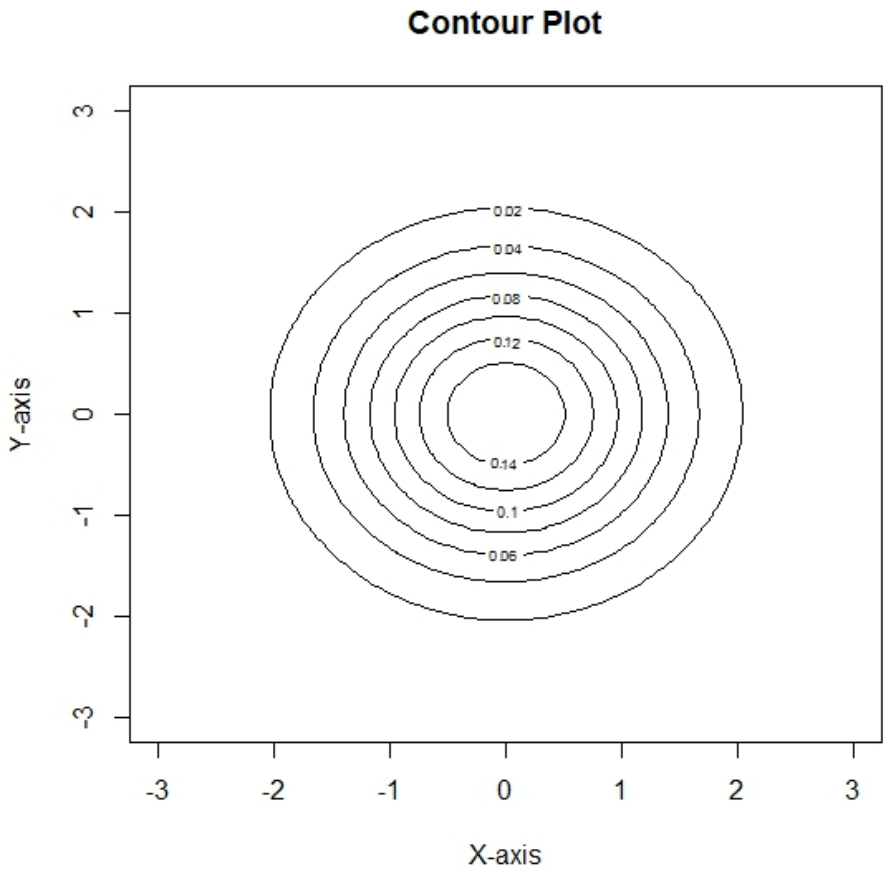
where n is the number of data points, h_X and h_Y are the bandwidth parameters for X and Y respectively, and $K(\cdot)$ is the kernel function, typically chosen to be the Gaussian kernel:

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}.$$

Contour plots - independent normals

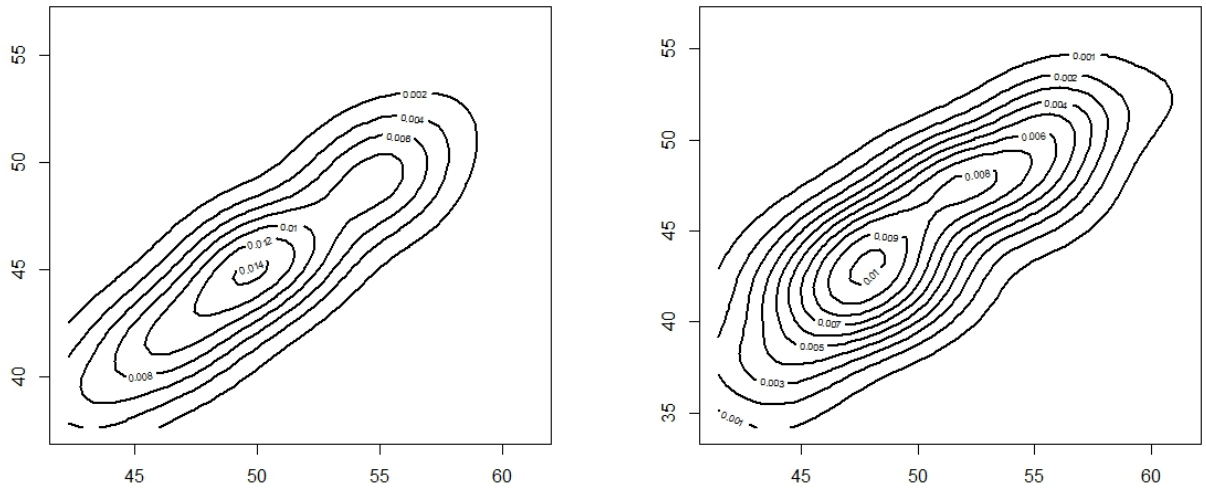
```
grid_x <- seq(-3, 3, length.out = 100)
grid_y <- seq(-3, 3, length.out = 100)
grid <- expand.grid(x = grid_x, y = grid_y)
grid$z <- with(grid, dnorm(x) * dnorm(y))

contour(grid_x, grid_y, matrix(grid$z, nrow = 100))
```

Contour plots - independent normals



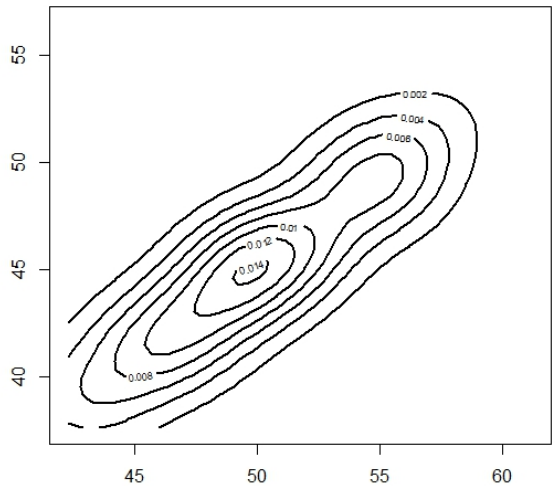
Contour plots - original vs. anonymized I

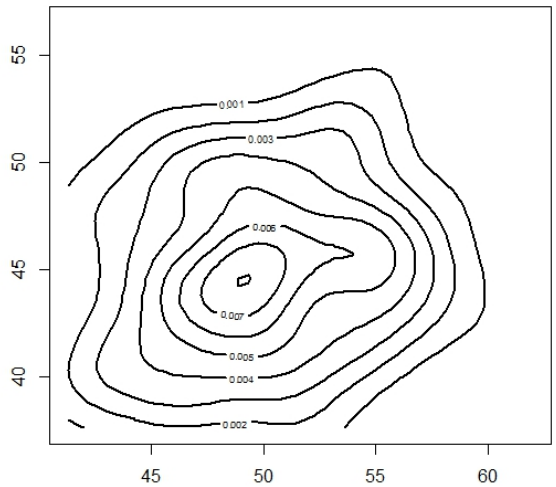
```
cont.orig <- kde2d(X1, X2, n = 50);
cont.anon <- kde2d(Y1, Y2, n = 50)
contour(cont.orig, lwd = 2); contour(cont.anon, lwd = 2)
```



Contour plots - original vs. anonymized II

```
cont.orig <- kde2d(X1, X2, n = 50);
cont.anon <- kde2d(U1, U2, n = 50)
contour(cont.orig, lwd = 2); contour(cont.anon, lwd = 2)
```





Propensity Score Analysis

This method calculates the probability that a data point is anonymized, helping to evaluate the indistinguishability of the original and anonymized data.
The propensity score is the conditional probability that a data point is anonymized given the observed and anonymized datasets. The process involves first combining the anonymized and original datasets and then adding an indicator variable to each data point to denote whether it is anonymized or not. This indicator variable is then used as the response to fit a model, such as logistic regression (logit) or classification trees (CART), with all other observed variables predicting the indicator. The fitted model provides predicted probabilities for each data point of being anonymized or original, which are the propensity scores. These scores represent the likelihood that each data point is anonymized on the basis of its characteristics, helping to measure the similarity between anonymized and observed data. The idea is that if the anonymized data is similar to observed data, it should be difficult to distinguish between them by propensity score.

Propensity Score Analysis

- n is the number of observed units;
- m is the number of anonymized units;
- $N = n + m$ is the total number of observations;
- $c = \frac{m}{N}$ is the proportion of anonymized units;
- t is an indicator variable that distinguishes between anonymized and observed data. It is added to the combined dataset to indicate which records are anonymized ($t = 1$) and which are observed ($t = 0$).

Propensity Score Analysis

Two typical approaches to model propensity scores:

- Logistic regression (Logit): Uses the glm function to fit a logistic regression model $t = f(X)$. The function $f(X)$ represents the logistic regression model, where X is the vector of features and $f(X)$ gives the probability that an observation is anonymized.
- Classification and Regression Trees (CART): Uses either the rpart or ctree function to fit a classification tree $t = f(X)$. The function $f(X)$ represents the classification tree, where X is the vector of characteristics, and $f(X)$ gives the probability that an observation is anonymized.

Propensity Score Analysis

Then, $\hat{p}_j$ is the propensity score for the jth observation, which is the predicted probability using model $f(x)$ introduced above. If propensity scores are close to c, then the assignment of these scores is essentially random. Then it indicates that it is difficult to distinguish between anonymized and original data. In turn, it means good data utility.

Propensity Score Analysis

We have the following measures based on the propensity score $\hat{p}(x_j)$.

- pMSE (**Propensity Mean Squared Error**)
  - pMSE measures the mean squared error of propensity scores, lower pMSE indicates better similarity between anonymized and observed data. The package does not use a test statistic that returns a p-value for pMSE.
  - Formula:

$$\text{pMSE} = \frac{1}{N}\sum_{j=1}^{N}(\hat{p}_j - c)^2.$$

  - A lower pMSE value indicates that the propensity scores are very close to c, suggesting that it is difficult to distinguish between anonymized and original data.

Propensity Score Analysis

- **Kolmogorov-Smirnov Statistic**
  - The Kolmogorov-Smirnov (KS) statistic measures the maximum difference between the empirical cumulative distribution functions of the propensity scores for observed and anonymized data. It is a statistics designed to test if two independent samples come from the same distribution. The package uses the KS test, which returns a p-value.
  - Formula: Denote $\hat{p}_{(x_j,\text{obs})}$ and $\hat{p}_{(x_j,\text{anon})}$ the propensity score for data point from observed and anonymized dataset respectively, $F_{\text{obs}}(x)$ and $F_{\text{syn}}(x)$ are the empirical cumulative distribution functions of the propensity scores for the observed and anonymized data, where

$$F_{\text{obs}}(x) = \frac{1}{n}\sum_j 1\{\hat{p}_{(x_j,\text{obs})} \leq x\} \text{ and } F_{\text{syn}}(x) = \frac{1}{m}\sum_j 1\{\hat{p}_{(x_j,\text{anon})} \leq x\},$$

for $x \in (0,1)$. We have

$$\text{KS} = \sup_x |F_{\text{obs}}(x) - F_{\text{anon}}(x)|.$$

Propensity Score Analysis

- **Wilcoxon Statistic**
  - The Wilcoxon rank-sum test is a nonparametric equivalent of the paired t-test. It makes no assumption about the distributions of the original population themselves, but it does assume that the distributions of the differences are at least symmetric.
  - Formula:

$$U = \sum_j \text{Rank}(\hat{p}_j^{\text{obs}}) - \frac{n(n+1)}{2}.$$

    where $\hat{p}_j^{\text{obs}}$ are the propensity scores for the observed data, and n is the number of observed units.

Example 3

Based on Example **??** now assume that we have 3 random variables: age ($X_{1,\text{obs}}$), income ($X_{2,\text{obs}}$), and sex ($X_{3,\text{obs}}$). The visit sequence for this example involves first generating age, then generating income based on age, and finally generating gender based on age and income. The basic implementation of the package is as follows.

1. Create $X_{1,\text{syn}}$ by bootstrapping $X_{1,\text{obs}}$.
2. Fit the linear model $X_{2,\text{obs}} = \beta X_{1,\text{obs}} + Z$. As a result, obtain the estimate $\hat{\beta}$.
3. Simulate $X_{2,\text{syn}}$ using $X_{2,\text{syn}} = \hat{\beta} X_{1,\text{syn}} + \tilde{Z}$, where $\tilde{Z}$ is another random noise.
4. Centering the data involves subtracting the mean of each predictor variable from the observed and synthetic values. The centered predictor variables are defined as follows:

$$X_{1,\text{obs, centered}} = X_{1,\text{obs}} - \bar{X}_{1,\text{obs}},$$

$$X_{2,\text{obs, centered}} = X_{2,\text{obs}} - \bar{X}_{2,\text{obs}},$$