# Methods for Risk Evaluation #1

Jingchen (Monika) Hu

Vassar College

Data Confidentiality

# Outline

# Outline

# What we have covered

- Disclosure risk lecture:
  - disclosure arises given intruder's knowledge about certain individuals in a publicly available sample
  - combination of variables makes individuals unique

# What we have covered

- Disclosure risk lecture:
  - disclosure arises given intruder's knowledge about certain individuals in a publicly available sample
  - combination of variables makes individuals unique

- Bayesian synthesis models lectures (2):
  - joint modeling vs FCS
  - joint synthesis
    - ★ bivariate normal
  - sequential synthesis
    - ★ continuous: normal regression
    - ★ categorical: logistic regression, multinomial logistic regression, Dirichlet-multinomial conjugacy, probit regression
    - ★ count: poisson regression

# What we have covered cont'd

- Methods for utility evaluation lectures (2):
  - global utility
    - ★ propensity score measure
    - ★ cluster analysis measure
    - ★ empirical CDF measure
  - analysis-specific utility
    - ★ examples of categorical variables and geographic variables
    - ★ combining rules for partially synthetic data
    - ★ combining rules for fully synthetic data
    - ★ interval overlap utility measure

# Overview

- Now, we can focus on disclosure risk evaluation of synthetic dataset
- Two types of disclosure risks:
  - identification disclosure
    - ★ 3 summaries: expected match risk, true match rate, false match rate
    - ★ demo with categorical data
    - ★ brainstorm and demo with continuous data
  - attribute disclosure
    - ★ examples of application-specific attribute disclosure risk

# Outline

# 3 summaries of identification disclosure risks

- the expected match risk
- the true match rate
- the false match rate

# 3 summaries of identification disclosure risks

- the expected match risk
- the true match rate
- the false match rate

Each is summarizing one aspect of identification disclosure risks

# Preliminaries

- $c_i$: the number of records with the highest match probability for the target record $i$.
  1. the records with the highest match probability for record $i$ are a subset of all the records sharing the same known information by the intruder.
  2. e.g. for categorical variable(s), we can consider all records in the same known pattern with record $i$.
  3. e.g. for continuous variable(s), we can consider all records within a certain distance from record $i$.
- $T_i$: if the true match is among the $c_i$ units, $T_i = 1$; otherwise $T_i = 0$.

# Preliminaries cont'd

- $K_i$: if the true match is the unique match (i.e. $c_i T_i = 1$), $K_i = 1$; otherwise $K_i = 0$.
- $F_i$: if there is a unique match but it is not the true match (i.e. $c_i(1 - T_i) = 1$), $F_i = 1$; otherwise $F_i = 0$.
- $N$: the total number of target records; typically $N = n$, the number of records in the sample.
- $s$: the number of uniquely matched records (i.e. $\sum_{i=1}^{n} c_i = 1$).

# The expected match risk

- On average how likely it is to find the correct match for each record, and for the sample as a whole

$$\sum_{i=1}^{n} \frac{T_i}{c_i} \tag{1}$$

# The expected match risk

- On average how likely it is to find the correct match for each record, and for the sample as a whole

$$\sum_{i=1}^{n} \frac{T_i}{c_i} \tag{1}$$

- For record $i$, $T_i$ indicates whether the true match is among the $c_i$ matched records
- When $T_i = 1$ and $c_i > 1$, the ratio $\frac{T_i}{c_i}$ refers to the probability of randomly guessing which of the $c_i$ matched records is the true match
- When $T_i = 0$, no matter how small $c_i$ is (e.g. $c_i = 1$ indicates only one matched record for record $i$), there is a 0 probability of guessing the identity of record $i$ correctly
- Note that when $c_i = 0$, we set $\frac{T_i}{c_i} = 0$.

# The expected match risk cont'd

$$\sum_{i=1}^{n} \frac{T_i}{c_i}$$

- Each $\frac{T_i}{c_i}$ is a record-level probability $\in [0, 1]$
- The sum $\sum_{i=1}^{n} \frac{T_i}{c_i}$ is a sample-level summary of the expected match risk, which is $\in [0, n]$
- The higher the expected match risk $\sum_{i=1}^{n} \frac{T_i}{c_i}$, the higher the identification disclosure risk for the sample, and vise versa

# The true match rate

- How large a percentange of true unique mathces exists

$$\sum_{i=1}^{n} \frac{K_i}{N} \tag{2}$$

# The true match rate

- How large a percentange of true unique mathces exists

$$\sum_{i=1}^{n} \frac{K_i}{N} \tag{2}$$

- For record $i$, $K_i = 1$ if the true match is the unique match (i.e. $c_i T_i = 1$)
- $N$ is the total number of target records (if we are evaluating the disclosure risk for every record in the sample, $N = n$)

# The true match rate

- How large a percentange of true unique mathces exists

$$\sum_{i=1}^{n} \frac{K_i}{N} \tag{2}$$

- For record $i$, $K_i = 1$ if the true match is the unique match (i.e. $c_i T_i = 1$)
- $N$ is the total number of target records (if we are evaluating the disclosure risk for every record in the sample, $N = n$)

- $\sum_{i=1}^{n} \frac{K_i}{N}$ is the percentage of true unique matches among the target records
- The higher the true match rate, the higher the identification disclosure risk for the sample, and vise versa

# The false match rate

- The percentage of unique matches to be false matches

$$\sum_{i=1}^{n} \frac{F_i}{s} \tag{3}$$

# The false match rate

- The percentage of unique matches to be false matches

$$\sum_{i=1}^{n} \frac{F_i}{s} \tag{3}$$

- For record $i$, $F_i = 1$ if there is a unique match but it is not the true match (i.e. $c_i(1 - T_i) = 1$)
- $s$ is the number of uniquely matched records (i.e. $\sum_{i=1}^{n} c_i = 1$)

# The false match rate

- The percentage of unique matches to be false matches

$$\sum_{i=1}^{n} \frac{F_i}{s} \tag{3}$$

- For record $i$, $F_i = 1$ if there is a unique match but it is not the true match (i.e. $c_i(1 - T_i) = 1$)
- $s$ is the number of uniquely matched records (i.e. $\sum_{i=1}^{n} c_i = 1$)

- $\sum_{i=1}^{n} \frac{F_i}{s}$ is the percentage of false matches among unique matches
- The lower the false match rate, the higher the identification disclosure risk for the sample, and vise versa

# Summary and discussion

- In sum, higher expected match risk, higher true match rate, and lower false match rate indicate higher identifiation disclosure risk for the sample
- When $m > 1$ synthetic datasets are generated, we can calculate the three summaries on each synthetic dataset, and take the average

# Summary and discussion

- In sum, higher expected match risk, higher true match rate, and lower false match rate indicate higher identifiation disclosure risk for the sample
- When $m > 1$ synthetic datasets are generated, we can calculate the three summaries on each synthetic dataset, and take the average

- Discussion questions:
  - what if the three summaries give inconsistent evaluation?
  - in what situation, one summary should be preferred over the others?

# Outline

# ACS sample information

| Variable | Information |
|---|---|
| SEX | 1 = male, 2 = female |
| RACE | 1 = White alone, 2 = Black or African American alone, 3 = American Indian alone, 4 = other, 5 = two or more races, 6 = Asian alone |
| MAR | 1 = married, 2 = widowed, 3 = divorced, 4 = separated, 5 = never married |
| LANX | 1 = speaks another language, 2 = speaks only English |
| WAOB | born in: 1 = US state, 2 = Puerto Rico and US island areas, oceania and at sea, 3 = Latin America, 4 = Asia, 5 = Europe, 6 = Africa, 7 = Northern America |
| DIS | 1 = has a disability, 2 = no disability |
| HICOV | 1 = has health insurance coverage, 2 = no coverage |
| MIG | 1 = live in the same house (non movers), 2 = move to outside US and Puerto Rico, 3 = move to different house in US or Puerto Rico |
| SCH | 1 = has not attended school in the last 3 months, 2 = in public school or college, 3 = in private school or college or home school |

# ACS sample information cont'd

- `ACSdata_org`: the original ACS sample
- `ACSdata_syn`: one synthetic ACS sample
  - four variables are synthesized: `LANX, WAOB, DIS, HICOV`
  - `m = 1` for illustration purpose

# ACS sample information cont'd

- `ACSdata_org`: the original ACS sample
- `ACSdata_syn`: one synthetic ACS sample
  - four variables are synthesized: `LANX, WAOB, DIS, HICOV`
  - `m = 1` for illustration purpose

- Known variables: `SEX, RACE, MAR`
- Goal: use this information to identify records in `ACSdata_syn`, obtain the 3 summaries

```
ACSdata_org <- read.csv(file = "ACSdata_org.csv")
ACSdata_syn <- read.csv(file = "ACSdata_syn.csv")
```

# Step 1: calculate key quantities

```
CalculateKeyQuantities <- function(origdata, syndata, known.vars, syn.vars, n){
  origdata <- origdata
  syndata <- syndata
  n <- n

  c_vector <- rep(NA, n)
  T_vector <- rep(NA, n)

  for (i in 1:n){
    match <- (eval(parse(text=paste("origdata$",syn.vars,"[i]==
                                    syndata$",syn.vars,sep="",collapse="&")))&
                eval(parse(text=paste("origdata$",known.vars,"[i]==
                                      syndata$",known.vars,sep="",collapse="&")))
    match.prob <- ifelse(match, 1/sum(match), 0)

    if (max(match.prob) > 0){
      c_vector[i] <- length(match.prob[match.prob == max(match.prob)])
    }
    else
      c_vector[i] <- 0
      T_vector[i] <- is.element(i, rownames(origdata)[match.prob == max(match.prob)]
  }
```

# Step 1: calculate key quantities cont'd

```
  K_vector <- (c_vector * T_vector == 1)
  F_vector <- (c_vector * (1 - T_vector) == 1)
  s <- length(c_vector[c_vector == 1 & is.na(c_vector) == FALSE])

  res_r <- list(c_vector = c_vector,
                T_vector = T_vector,
                K_vector = K_vector,
                F_vector = F_vector,
                s = s
  )
  return(res_r)
}
```

# Step 1: calculate key quantities cont'd

- four synthesized variables: `LANX`, `WAOB`, `DIS`, `HICOV`, assigned to `syn.vars`
- three known variables: `SEX`, `RACE`, `MAR`, assigned to `known.vars`

```
known.vars <- c("SEX", "RACE", "MAR")
syn.vars <- c("LANX", "WAOB", "DIS", "HICOV")
n <- dim(ACSdata_org)[1]

KeyQuantities <- CalculateKeyQuantities(ACSdata_org, ACSdata_syn,
                                        known.vars, syn.vars, n)
```

# Step 2: calculate 3 summary measures

```
IdentificationRisk <- function(c_vector, T_vector, K_vector, F_vector, s, N){

  nonzero_c_index <- which(c_vector > 0)

  exp_match_risk <- sum(1/c_vector[nonzero_c_index]*T_vector[nonzero_c_index])
  true_match_rate <- sum(na.omit(K_vector))/N
  false_match_rate <- sum(na.omit(F_vector))/s

  res_r <- list(exp_match_risk = exp_match_risk,
                true_match_rate = true_match_rate,
                false_match_rate = false_match_rate
  )
  return(res_r)
}
```

# Step 2: calculate 3 summary measures cont'd

- each record is a target, therefore `N = n`

```
c_vector <- KeyQuantities[["c_vector"]]
T_vector <- KeyQuantities[["T_vector"]]
K_vector <- KeyQuantities[["K_vector"]]
F_vector <- KeyQuantities[["F_vector"]]
s <- KeyQuantities[["s"]]
N <- n

ThreeSummaries <- IdentificationRisk(c_vector, T_vector, K_vector, F_vector, s, N)
```

# Step 2: calculate 3 summary measures cont'd

```
ThreeSummaries[["exp_match_risk"]]
```

```
## [1] 41.36863
```

```
ThreeSummaries[["true_match_rate"]]
```

```
## [1] 5e-04
```

```
ThreeSummaries[["false_match_rate"]]
```

```
## [1] 0.974359
```

# Results and discussion

- The 41.37 expected match risk: $\frac{41.37}{10000} = 0.000042$ probability on average for each record to be correctly identified

# Results and discussion

- The 41.37 expected match risk: $\frac{41.37}{10000} = 0.000042$ probability on average for each record to be correctly identified
- The 0.0005 true match rate: $0.0005 \times 10000 = 5$ records are correct unique matches

# Results and discussion

- The 41.37 expected match risk: $\frac{41.37}{10000} = 0.000042$ probability on average for each record to be correctly identified
- The 0.0005 true match rate: $0.0005 \times 10000 = 5$ records are correct unique matches
- The 0.97 false match rate: among the 195 (the value of $s$) unique matches, 190 are false matches, i.e. they are not the true matches

# Results and discussion

- The 41.37 expected match risk: $\frac{41.37}{10000} = 0.000042$ probability on average for each record to be correctly identified
- The 0.0005 true match rate: $0.0005 \times 10000 = 5$ records are correct unique matches
- The 0.97 false match rate: among the 195 (the value of $s$) unique matches, 190 are false matches, i.e. they are not the true matches
- Overall, the identification disclosure risks for the synthetic ACS sample seem very low, indicating a high level of confidentiality protection of the synthetic ACS data

# Results and discussion cont'd

- A good practice to write functions to calculate various quantities
- When `m > 1`
  - create `c_vector, T_vector, K_vector, F_vector` as matrices $(n - by - m)$
  - create `s` as a vector of length $m$
  - add nested loops when necessary
  - create `exp_match_risk, true_match_rate, false_match_rate` as vectors of length $m$
  - `syndata` is a list