analysis_3_24

Kevin Ros

3/17/2020

```
ACSdata_org <- read.csv(file = "ACSdata_org.csv")
ACSdata_syn <- read.csv(file = "ACSdata_syn.csv")
ACSdata_syn2 <- read.csv(file = "ACSdata_syn2.csv")
ACSdata_syn3 <- read.csv(file = "ACSdata_syn3.csv")
```

```
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)])
  }

  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)
}
```

```
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)
KeyQuantities2 <- CalculateKeyQuantities(ACSdata_org, ACSdata_syn2,
                                        known.vars, syn.vars, n)
KeyQuantities3 <- CalculateKeyQuantities(ACSdata_org, ACSdata_syn3,
```

```
                                        known.vars, syn.vars, n)
quantities = list(KeyQuantities, KeyQuantities2, KeyQuantities3)
```

```
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)
}
```

```
avg_exp = 0
avg_true = 0
avg_false = 0
for(q in quantities){
  c_vector <- q[["c_vector"]]
  T_vector <- q[["T_vector"]]
  K_vector <- q[["K_vector"]]
  F_vector <- q[["F_vector"]]
  s <- q[["s"]]
  N <- n
  ThreeSummaries <- IdentificationRisk(c_vector, T_vector, K_vector, F_vector, s, N)
  avg_exp = avg_exp + ThreeSummaries[["exp_match_risk"]]
  avg_true = avg_true + ThreeSummaries[["true_match_rate"]]
  avg_false = avg_false + ThreeSummaries[["false_match_rate"]]
}
```

```
avg_exp / 3
```

```
## [1] 41.46743
```

```
avg_true / 3
```

```
## [1] 0.0005666667
```

```
avg_false / 3
```

```
## [1] 0.9638026
```