Testing different imputation methods on PUMS (MAR) - MICE-CART

```r
# load dataset: df
load('../Datasets/ordinalPUMS.Rdata')

# take 10,000 samples: df
set.seed(0)
n = 10000
sample <- sample(nrow(df), size = 10000)
df <- df[sample,]

# create MCAR scneario with 30% chance of missing: df_observed
missing_prob = 0.3
df_observed <- df
missing_col = c(1,3,7,9,10,11)

# Make VEH and WKL MCAR
missing_col_MCAR = c(1,10)
for (col in missing_col_MCAR) {
  missing_ind <- rbernoulli(n,p = missing_prob)
  df_observed[missing_ind, col] <- NA
}

# Make the rest MAR
numeric_df = sapply(df, as.numeric)
normalized_df = t(t(numeric_df-1)/(apply(numeric_df, MARGIN = 2, FUN = max)-1))
missing_col_MAR = c(3,7,9,11)
fully_observed_col = c(2,4,5,6,8)
beta_NP = c(0, 15, -12, 14, 0)
beta0_NP = -1
beta_SCHL = c(2, -12, 2.5, 0, 0)
beta0_SCHL = -1
beta_AGEP = c(4, 5, 0, -12, 0)
beta0_AGEP = -1
beta_PINCP = c(-12, 0, 4, 4, 0)
beta0_PINCP = -1

# missing probability for NP
prob_NP = apply(t(t(normalized_df[, fully_observed_col])*beta_NP), MARGIN = 1, sum)+beta0_NP
prob_NP = exp(prob_NP)/(exp(prob_NP)+1)
indicator = rbernoulli(n, p = prob_NP)
df_observed[indicator, missing_col_MAR[1]] <- NA

# missing probability for SCHL
prob_SCHL = apply(t(t(normalized_df[, fully_observed_col])*beta_SCHL), MARGIN = 1, sum) + beta0_SCHL
prob_SCHL = exp(prob_SCHL)/(exp(prob_SCHL)+1)
indicator = rbernoulli(n, p = prob_SCHL)
df_observed[indicator, missing_col_MAR[2]] <- NA
```

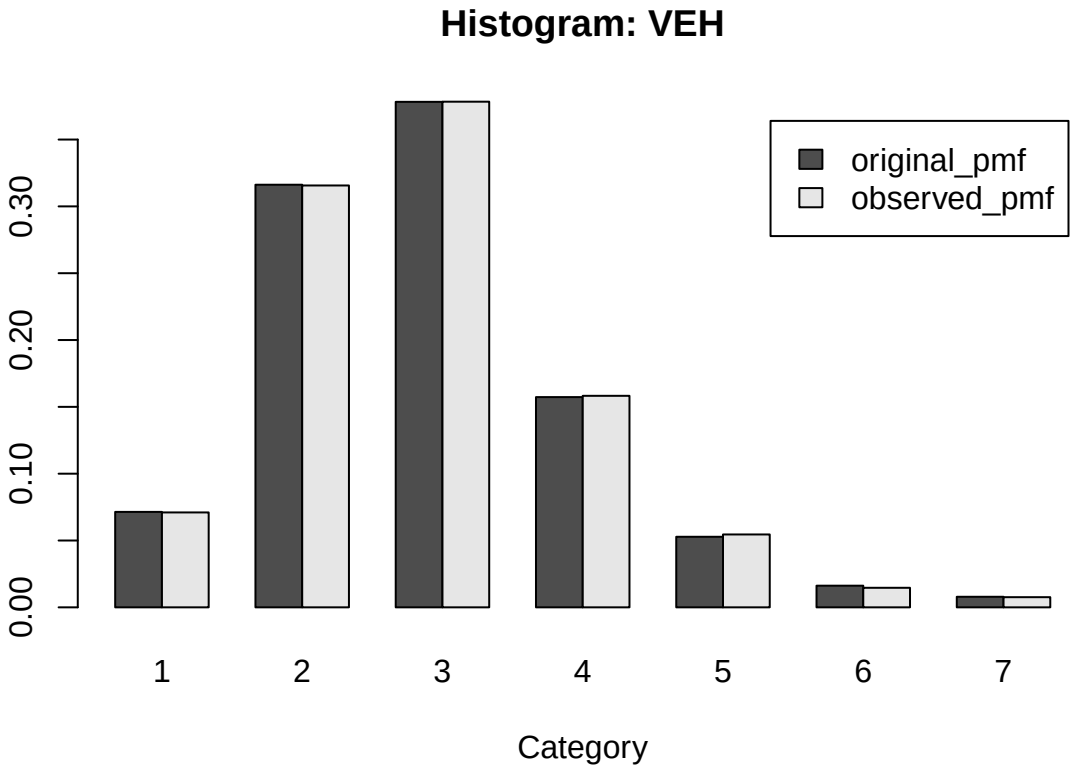
```r
# missing probability for AGEP
prob_AGEP = apply(t(t(normalized_df[, fully_observed_col])*beta_AGEP), MARGIN = 1, sum) + beta0_AGEP
prob_AGEP = exp(prob_AGEP)/(exp(prob_AGEP)+1)
indicator = rbernoulli(n, p = prob_AGEP)
df_observed[indicator, missing_col_MAR[3]] <- NA

# missing probability for PINCP
prob_PINCP = apply(t(t(normalized_df[, fully_observed_col])*beta_PINCP), MARGIN = 1, sum) + beta0_PINCP
prob_PINCP = exp(prob_PINCP)/(exp(prob_PINCP)+1)
indicator = rbernoulli(n, p = prob_PINCP)
df_observed[indicator, missing_col_MAR[4]] <- NA

# 30.61% missing
apply(is.na(df_observed), MARGIN = 2, mean)
```

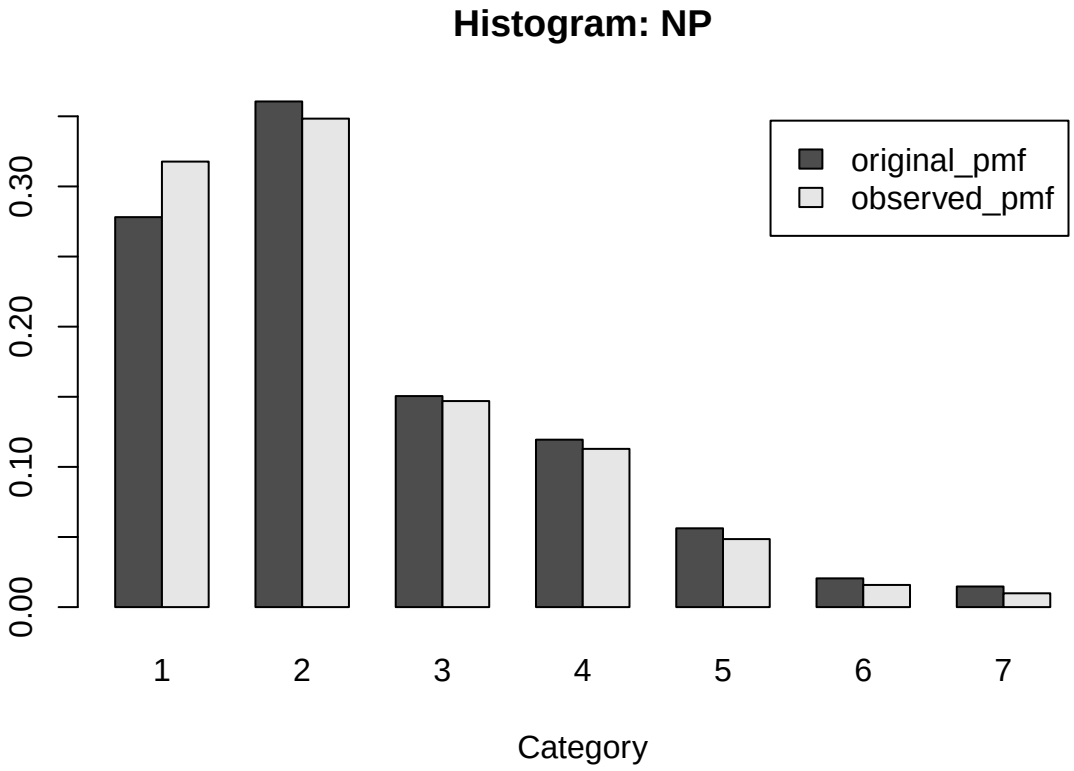
```
##    VEH     MV     NP   RMSP    ENG  MARHT   SCHL RACNUM   AGEP    WKL  PINCP
## 0.3030 0.0000 0.2786 0.0000 0.0000 0.0000 0.3159 0.0000 0.3273 0.3017 0.3101
```

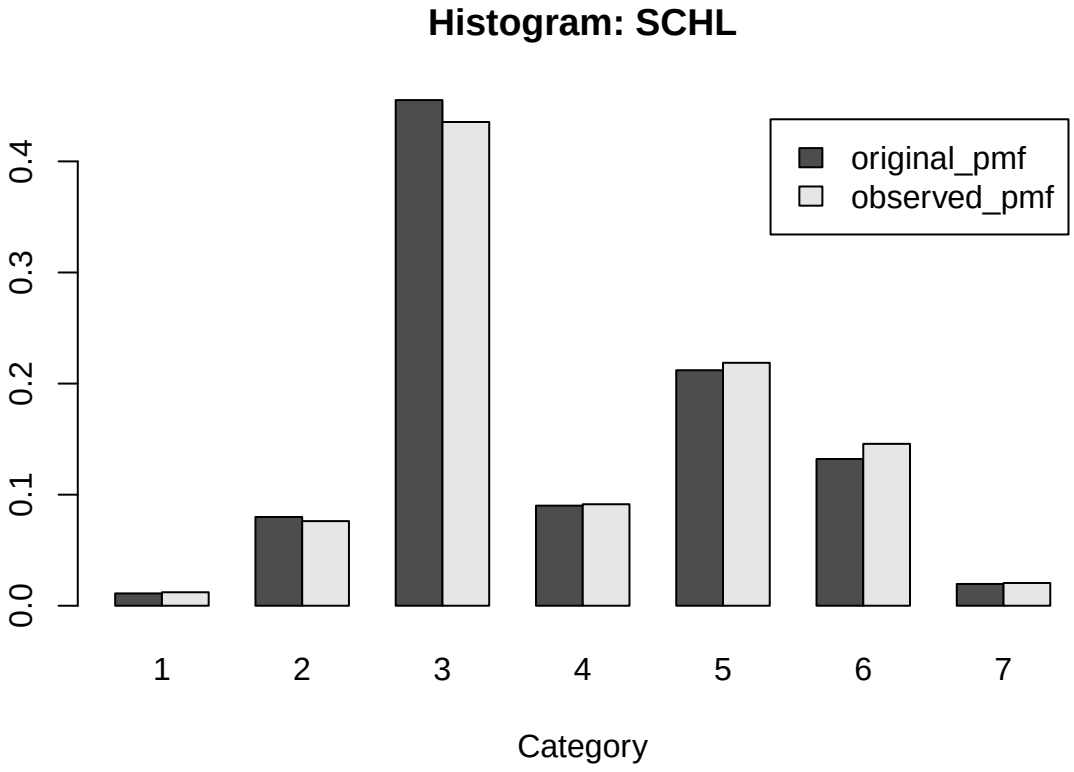
Histogram for univariate distribution

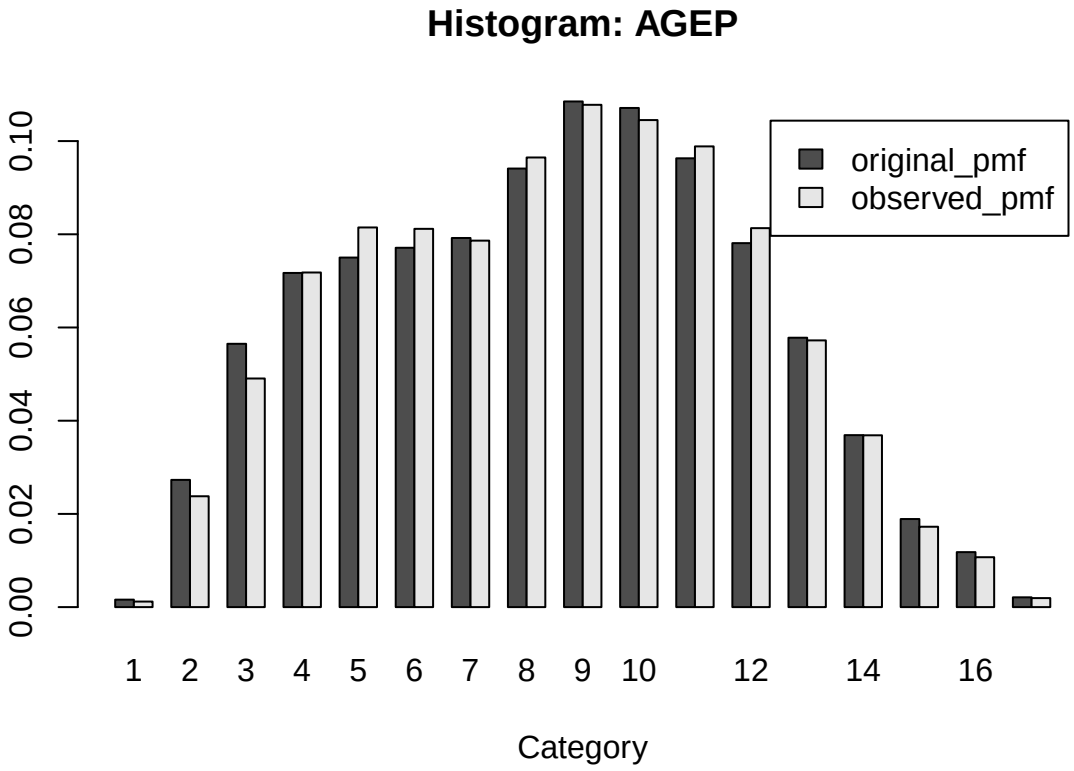


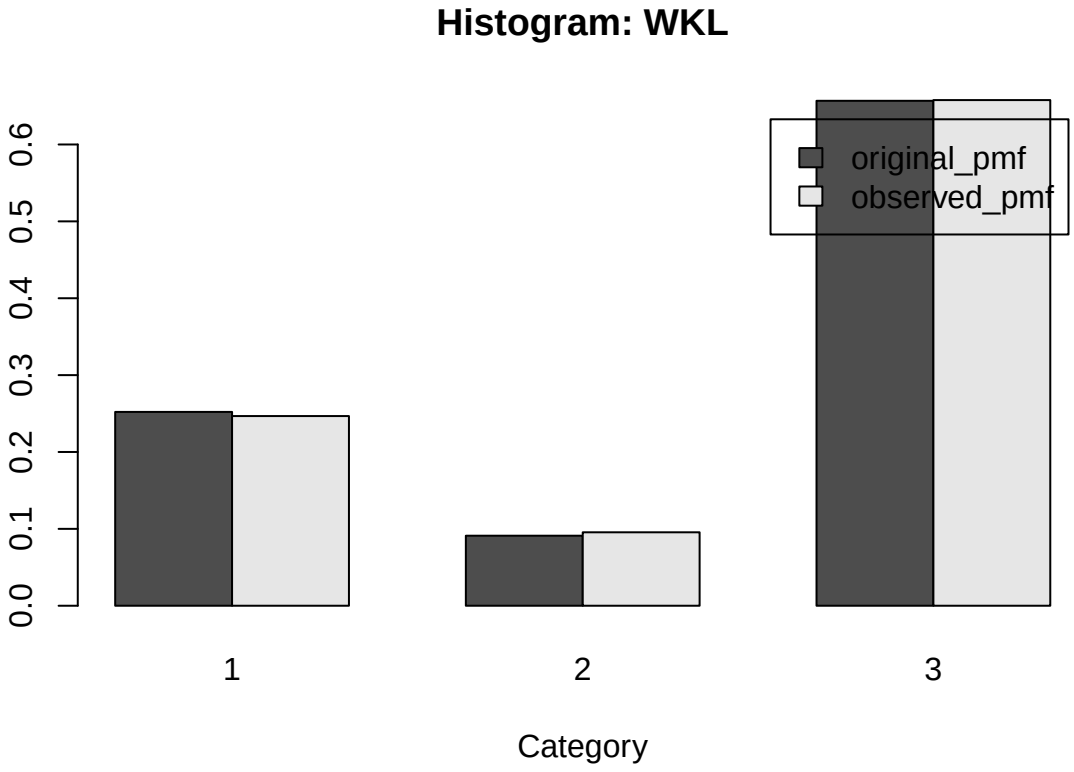
Assess bivariate joint distribution

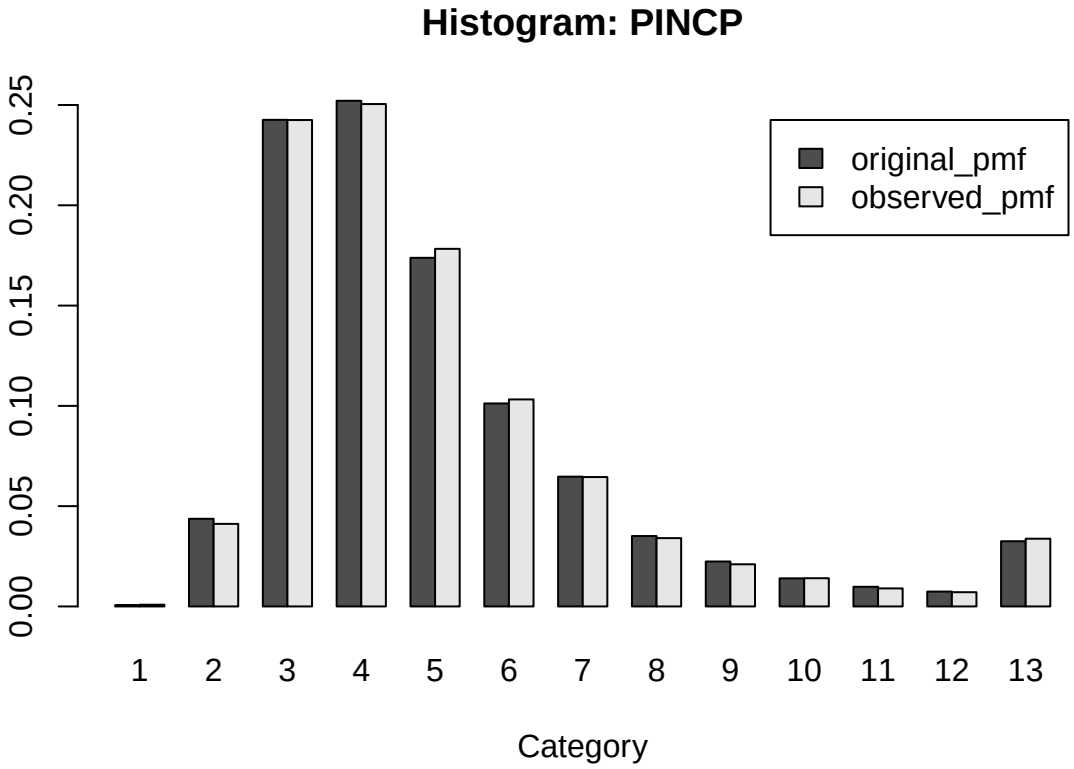
Assess trivariate joint distribution

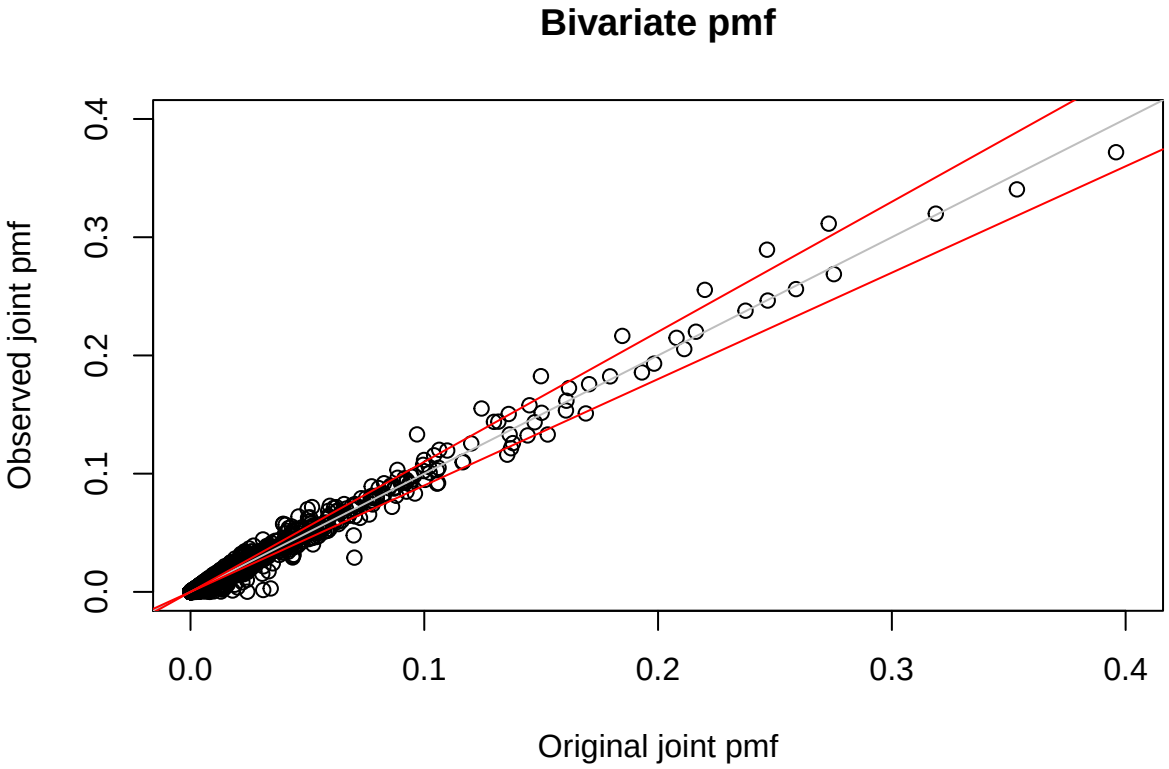

Histogram: NP
original_pmf
observed_pmf
0.00
0.10
0.20
0.30
1
2
3
4
5
6
7
Category

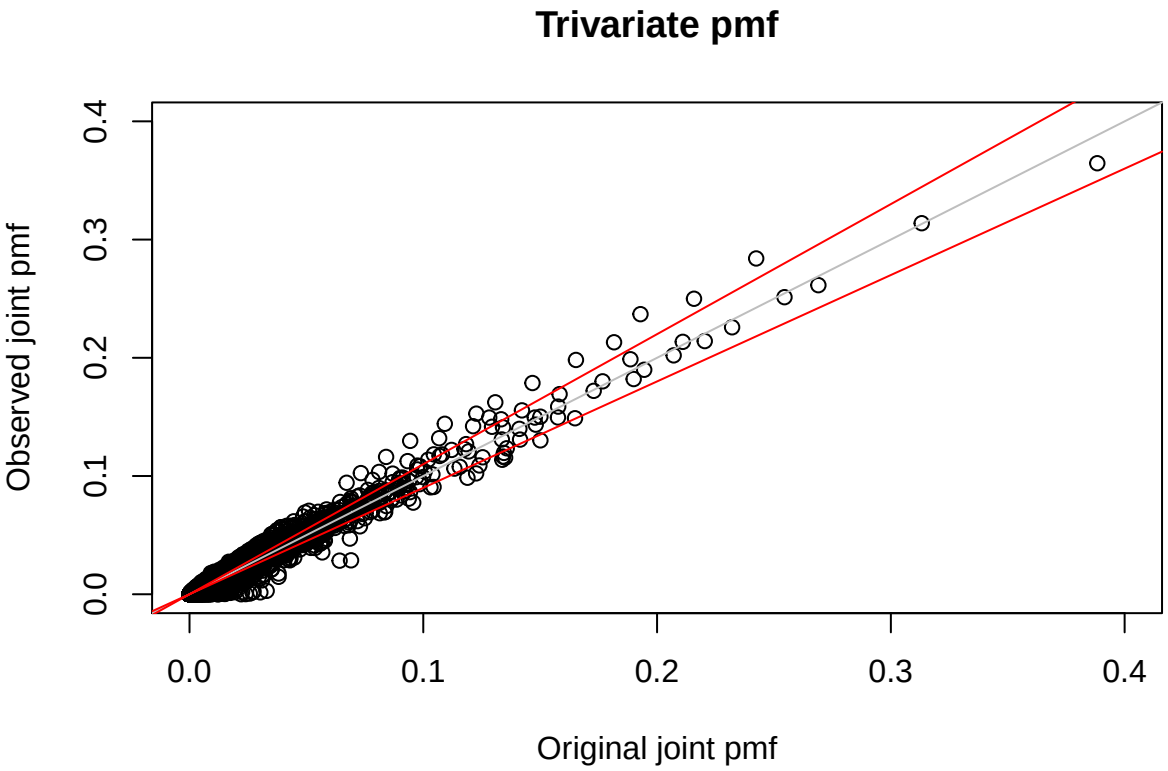

Histogram: SCHL
original_pmf
observed_pmf
0.0
0.1
0.2
0.3
0.4
1
2
3
4
5
6
7
Category


Histogram: AGEP
0.00
0.02
0.04
0.06
0.08
0.10
1
2
3
4
5
6
7
8
9
10
12
14
16
Category
original_pmf
observed_pmf


Histogram: WKL
0.0
0.1
0.2
0.3
0.4
0.5
0.6
1
2
3
Category
original_pmf
observed_pmf


Histogram: PINCP
original_pmf
observed_pmf
0.00
0.05
0.10
0.15
0.20
0.25
1
2
3
4
5
6
7
8
9
10
11
12
13
Category


Bivariate pmf
Observed joint pmf
0.0
0.1
0.2
0.3
0.4
Original joint pmf


Trivariate pmf
Observed joint pmf
Original joint pmf
0.0
0.1
0.2
0.3
0.4