Mixing Matrix Estimation

Caleb Easterly

The goal is to estimate the annual number of new partnerships between each group of people in the model. In this model, we have 2 sexes (male and female), three sexual identities (heterosexual, gay, and bisexual), and two sexual activity groups (high and low). That gives a total of 12 demographic groups.

In the Natsal data, we know the sex, self-reported sexual identity (sexID), and sexual activity level of the respondents, as well as the gender of the respondents' partners. The main problem is that the respondents they don't report (and/or don't know) the sexual orientation or sexual activity of their partner. So, we have the total number of partners that each group has with each gender, but we'll have to distribute these partnerships among sexual orientations and sexual activities.

Because there doesn't seem to be much information on how people mix regarding sexual orientation or sexual activity, we make the proportionate assumption, which amounts to assuming that people have no preference for sexual orientation or sexual activity - they simply choose among what is available to them.

In the paper, we use three model populations:

1. Heterosexual population, sexual activity stratification only
2. Het, gay/lesbian, bi population, sexual activity stratification only
3. Het, gay/lesbian, bi population, sexual activity and sexID stratification

We carry out the analysis three times, doing the same three steps each time:

1. Estimate the number of new partnerships with each sex
2. Distribute partnerships across sexual identities and sexual activity groups using the proportionality assumption
3. Balance the resulting distributed partnerships

Data prep

First, we read in the Natsal data and prepare it for analysis. We use the following R packages:

```r
library(dplyr)
library(reshape2)
library(ggplot2)
library(knitr)
library(kableExtra)
```

This is a function to pretty-print tables.

```r
format_table <- function(df) {
  df %>%
    knitr::kable(booktabs = TRUE) %>%
    kableExtra::kable_styling(bootstrap_options = "striped",
                              latex_options = c("scale_down", "striped"),
                              font_size = 8)
}
```

Relevant Variables in Natsal

The following variables in the Natsal data set are relevant to our analysis:

- rsex: respondent's sex
  - 1 is male
  - 2 is female
- sexid: sexual identity
  - 1: heterosexual / straight
  - 2: gay / lesbian
  - 3: bisexual
  - 4: other
  - 9: not answered
- everhet: ever had a heterosexual sexual partnership
  - 1 is yes
  - 0 is no
  - 9 is "unclassifiable"
- eversam: Ever had a same-sex partnership
  - 1 is yes
  - 2 is no
  - 9: NA
  - -1: not answered
- hetnonew: total number of new heterosexual partners in the past year
  - 0: if het1yr=0 | (het1yr=1 & hetnewp=2)
  - 1: if het1yr=1 & hetnewp=1
  - 99: not answered if missing
  - -1: not applicable if het1yr=-1
- samnonew: no. of new hom. sex partners, last year
  - 0 if sam1yr=0 | (sam1yr=1 & samnewp=2)
  - 1 if sam1yr=1 & samnewp=1
  - copy hnonewp if sam1yr>1 & sam1yr<995
  - 999: not answered if missing
  - -1: not applicable if sam1yr=-1
- totnewyr: no. of new het. & hom. sex partners, last year
  - compute hetnonew + samnonew if hetnonew>=0 & hetnonew<999 & samnonew>=0 & samnonew<999
  - else hetnonew if samnonew=-1 & hetnonew>=0 & hetnonew<999
  - else samnonew if hetnonew=-1 & samnonew>=0 & samnonew<999
  - 999: not answered if hetnonew=999 | samnonew=999
  - -1: not applicable if hetnonew =-1 | samnonew =-1

Code Variable Names

Throughout this document, we use the following nomenclature:

Name	Definition
`r_sex`	the sex of the survey respondent
`r_sexid`	the self-reported sexual identity of the survey respondent
`r_sexact`	the sexual activity group of the survey respondent, based on their total number of partners
`prop`	the proportion of the total population with the designated sex, sexid, and/or sexact. This may be suffixed by `r_rp` or `rp_r`
`rp_sex`	the sex of the sex partners (respondents' partners). This is reported by the survey respondents.
`rp_sexid`	the sexual identity of the sex partners of the respondents. This will be estimated using the proportionality assumption

Name	Definition
`rp_sexact`	the sexual activity group of the sex partners of the respondents. This will be estimated using the proportionality assumption.
`partners`	the per-person number of new sex partners that the `r` group reported with people of sex `rp_sex`
`n_partners`	the total number of new sex partners that the `r` group reported with people of sex `rp_sex`
`partners.r_rp`	After combining respondents and hypothetical respondents' partners, the per-person number of partners that `r` has with `rp`
`partners.rp_r`	The hypothetical per-person number of partnerships that `rp` has with `r`
`prop_to_r`	The proportion of all partnerships offered to `r` that come from `rp`
`d_partners`, `d_partners.r_rp`, `d_partners.rp_r`	The per-person number of partnerships, distributed over sexids and/or sex activity groups. Either as reported, calculated from `r` to `rp`, or calculated from `rp` to `r`, respectively.
`n_partners*`, `n_d_partners*`	The total number of partnerships, i.e., the product of the partnership measure and the proportion in that group
`corrected_r`, `corrected_rp`	Per-person number of partnerships corrected for balancing, from the perspective of `r` and `rp` groups, respectively.
`cnr`, `cnrp`	The total number of partnerships from `r` to `rp` and vice versa. Used to check that the balancing worked

Analysis

Load NATSAL data in R form:

```r
load("../old/sfceSO/natsal_R_df.rda")
```

We focus on a higher-risk age group: 20-year-olds to 35-year-olds.

```r
MIN_AGE <- 20
MAX_AGE <- 35

natsal_hr <- filter(natsal_R_df, dage >= MIN_AGE & dage <= MAX_AGE)
```

Next, we make new indicator variables, recode some variables to make their values more transparent, and select only relevant variables.

We make two other choices:

1. Exclude anyone who doesn't report a sexual identity of heterosexual, gay, or bisexual. A large percentage (over 99%) report one of these three sexual identities.

```r
natsal_filt <- natsal_hr %>%
  filter(sexid == 1 | sexid == 2 | sexid == 3)
nrow(natsal_filt) / nrow(natsal_hr)
```

```
## [1] 0.9940724
```

2. Remove missing partner data. For `hetnonew`, "99" indicates missingness and "-1" indicates "not applicable". For `samnonew`, "999" indicates missingness and "-1" is not applicable. Again, the missing data is a small proportion of the total (less than 3%).

```r
natsal_filt2 <- natsal_filt %>%
  filter(hetnonew < 99 & samnonew < 995 & hetnonew >= 0 & samnonew >= 0)
nrow(natsal_filt2)/nrow(natsal_filt)
```

```
## [1] 0.974859
```

Now, we recode the sexid to allow for easier interpretation.

```
recode_sexid <- function(sexid){
  new_sexid <- rep(NA, length(sexid))
  new_sexid[sexid == 1] <- 'het'
  new_sexid[sexid == 2] <- 'gay'
  new_sexid[sexid == 3] <- 'bi'
  return(new_sexid)
}

all_sex <- natsal_filt2 %>%
  # make new indicator for female, eversam and everhet
  mutate(
      eversam_ind = (eversam == 1),
      everhet_ind = (everhet == 1),
      r_sex = ifelse(rsex == 1, "m", "w"),
      r_sexid = recode_sexid(sexid)
  ) %>%
  select(r_sex, r_sexid, eversam_ind, everhet_ind, hetnonew, samnonew, totnewyr, total_wt)
```

In all, we end up with 6049 observations, which is 96.9080423% of the original respondents between 20 and 35 years old.

Next, we define the 'sexual activity group', based on whether someone reports having 0-1 (low risk) or 2+ (high risk) new sex partners in the past year.

```
def_sex_act_group <- function(totnewyr){
  lent <- length(totnewyr)
  ret <- rep(NA, lent)
  ret[totnewyr >= 0 & totnewyr <= 1] <- 'low'
  ret[totnewyr >= 2] <- 'high'
  return(ret)
}

all_sex$r_sexact <- def_sex_act_group(all_sex$totnewyr)
```

These are the first few rows of the final cleaned data:

```
format_table(head(all_sex))
```

r_sex	r_sexid	eversam_ind	everhet_ind	hetnonew	samnonew	totnewyr	total_wt	r_sexact
m	het	FALSE	TRUE	1	0	1	0.4605145	low
w	het	FALSE	TRUE	0	0	0	1.7889542	low
m	het	FALSE	TRUE	0	0	0	1.3815434	low
w	het	FALSE	TRUE	1	0	1	0.7390113	low
m	het	TRUE	TRUE	3	0	3	0.9925568	high
m	het	FALSE	FALSE	0	0	0	0.5595336	low

Heterosexual population

This mixing matrix is for the heterosexual-only model. There are two sexes, one sexual orientation, and two sexual activity groups, so the matrix will be 4 rows by 4 columns. As stated above, the three steps are: estimate, distribute, and balance.

Estimate the number of new partnerships with each sex

This is a function that we'll need to use to get the proportion of the population in each group. This is important because we calculate the "total number of partnerships" as the per-person number of partnerships in a certain group times the proportion of the population that group. For example, if high-SA women had 2 partnerships per year on average and 25% of the population was high-SA women, then they would have $2 \times 0.25 = 0.5$ total partnerships per year. This number is called `n_partners` in the code below. The reason that we use the proportion, rather than some total number of people, is that the size of the population doesn't matter for the dynamic model. To get the absolute numbers for a population of, say, 100,000 people, we can just multiply the total number of partnerships by 100,000. That is, using the example above, high-SA women would "offer" 50,000 partnerships per year. Also note that the function uses the survey weights, rather than the number of respondents.

```
# the q stands for "query"
# denom stands for denominator
get_proportion <- function(natsal,
                           q_sex = c("m", "w"),
                           q_sexid = c("het", "gay", "bi"),
                           q_sexact = c("high", "low"),
                           denom_sex = c("m", "w"),
                           denom_sexid = c("het", "gay", "bi"),
                           denom_sexact = c("high", "low")){
    denom <- sum(natsal$total_wt[natsal$r_sex %in% denom_sex &
                                 natsal$r_sexid %in% denom_sexid &
                                 natsal$r_sexact %in% denom_sexact])
    qpop <- sum(natsal[natsal$r_sex %in% q_sex &
                          natsal$r_sexid %in% q_sexid &
                          natsal$r_sexact %in% q_sexact, "total_wt"])
    qpop / denom
}
```

Let's test this function. What proportion of heterosexual men have high and low sexual activity?

```
# high
get_proportion(all_sex, q_sex = "m", q_sexact = "high", q_sexid = "het",
               denom_sex = "m", denom_sexid = "het")
```

```
## [1] 0.1439956
```

```
# low
get_proportion(all_sex, q_sex = "m", q_sexact = "low", q_sexid = "het",
               denom_sex = "m", denom_sexid = "het")
```

```
## [1] 0.8560044
```

They add up to 1, as should be the case.

Now, we calculate the average number of new partners per-person with each sex (`rp_sex`), as well as the total number of new partners (`n_partners`).

```
het_sexact_rep <- all_sex %>%
    filter(r_sexid == "het") %>%
    group_by(r_sex, r_sexact, r_sexid) %>%
    summarise(partners = weighted.mean(hetnonew, w = total_wt)) %>%
    mutate(rp_sex = ifelse(r_sex == "m", 'w', 'm')) %>%
    rowwise() %>%
    mutate(prop = get_proportion(all_sex,
                                 q_sex = r_sex,
                                 q_sexact = r_sexact,
                                 q_sexid = "het",
                                 denom_sexid = "het"),
           n_partners = partners * prop) %>%
    ungroup()

format_table(het_sexact_rep)
```

r_sex	r_sexact	r_sexid	partners	rp_sex	prop	n_partners
m	high	het	4.0068490	w	0.0722716	0.2895815
m	low	het	0.2137853	w	0.4296301	0.0918486
w	high	het	3.2675397	m	0.0449038	0.1467251
w	low	het	0.1723069	m	0.4531944	0.0780885

So, high activity heterosexual men report an average of about 4 new partners per year, and they make up about 7.2% of the total population (note this is roughly half of the proportion of *men* that have high sexual activity). On average, then, the group of high activity men offers a total of about 0.29 partners per year.

Next, let's distribute the partnerships across activity levels.

Distribute partnerships

Since the respondent's don't know their partners' activity levels, we make the proportionality assumption and then estimate the proportion of partnerships from each activity level.

This amounts to

$$\Pr(S_{rp} = s', A_{rp} = a' | S_r = s, A_r = a) = \frac{N_{s'a'}\beta_{s'a's}}{\sum_i N_{s'i}\beta_{k'is}}$$

where:

- S_r and A_r are the sex and SA of the respondent,
- S_{rp} and A_{rp} are the sex and SA of the respondent's partner,
- N_{ka} is the proportion of people with sex k who have sexual activity a,
- $\beta_{kak'}$ is the number of partnerships that people of sex s' with sexual activity a' reported with sex s,
- and the denominator is the total number of partnerships offered to sex s from all sexual activity groups (note that, because this is a heterosexual model, the sex is assumed to be the opposite sex - i.e., s')

We can do this using dplyr and piping. Basically, for each combination of `r_sex` and `rp_sex`, we take the total number of partnerships 'offered' by each sex and sexual activity group, and divide it by the total number of partnerships offered by that sex. Then, we define `rp_sexid = r_sexid`, so the proportion `prop_to_r` represents the proportion of partnerships from `rp_sex` offered to `r_sex` that come from people with sex `rp_sex` and SA group `rp_sexact`.

```
het_sexact_offered_dist <-  het_sexact_rep %>%
    group_by(rp_sex, r_sex) %>%
    mutate(prop_to_r = n_partners / sum(n_partners)) %>%
    select(r_sex,
           rp_sex,
           rp_sexact = r_sexact,
           prop_to_r) %>%
    ungroup()
format_table(het_sexact_offered_dist)
```

r_sex	rp_sex	rp_sexact	prop_to_r
m	w	high	0.7591994
m	w	low	0.2408006
w	m	high	0.6526522
w	m	low	0.3473478

As an example, the above data shows that women with high sexual actvity account for 75.9% of all partnerships "offered" to men, while women with low sexual activity account for 24.1%.

Now, we distribute the partnerships across the groups. To do this, we join the proportion dataframe with the survey dataframe. This join aligns the `prop_to_r` and `partners` column, so they can be multiplied.

```
joined_dfs <- left_join(het_sexact_offered_dist, het_sexact_rep, by = c("r_sex", "rp_sex"))
format_table(joined_dfs)
```

r_sex	rp_sex	rp_sexact	prop_to_r	r_sexact	r_sexid	partners	prop	n_partners
m	w	high	0.7591994	high	het	4.0068490	0.0722716	0.2895815
m	w	high	0.7591994	low	het	0.2137853	0.4296301	0.0918486
m	w	low	0.2408006	high	het	4.0068490	0.0722716	0.2895815
m	w	low	0.2408006	low	het	0.2137853	0.4296301	0.0918486
w	m	high	0.6526522	high	het	3.2675397	0.0449038	0.1467251
w	m	high	0.6526522	low	het	0.1723069	0.4531944	0.0780885
w	m	low	0.3473478	high	het	3.2675397	0.0449038	0.1467251
w	m	low	0.3473478	low	het	0.1723069	0.4531944	0.0780885

The multiplication gives us the distributed parterships. Then, we define a single variable that describes the two demographic variables (sex and SA group):

```
partner_dist_het_sexact <- joined_dfs %>%
  mutate(d_partners = partners * prop_to_r,
         r_demo = paste(r_sex, r_sexact, sep="_"),
         rp_demo = paste(rp_sex, rp_sexact, sep="_")) %>%
  select(r_sex, r_sexact, r_sexid, rp_sex, rp_sexact, r_demo, rp_demo, d_partners, prop)
format_table(partner_dist_het_sexact)
```

r_sex	r_sexact	r_sexid	rp_sex	rp_sexact	r_demo	rp_demo	d_partners	prop
m	high	het	w	high	m_high	w_high	3.0419974	0.0722716
m	low	het	w	high	m_low	w_high	0.1623057	0.4296301
m	high	het	w	low	m_high	w_low	0.9648517	0.0722716
m	low	het	w	low	m_low	w_low	0.0514796	0.4296301
w	high	het	m	high	w_high	m_high	2.1325671	0.0449038
w	low	het	m	high	w_low	m_high	0.1124565	0.4531944
w	high	het	m	low	w_high	m_low	1.1349726	0.0449038
w	low	het	m	low	w_low	m_low	0.0598504	0.4531944

So, men with high sexual activity reported 4.01 partners. In the distributed table, these ~4 partners are distributed across women with high sexual activity and women with low sexual activity. Note that the proportions are roughly 75% and 25%, which is the proportion of partnerships that high SA women and low SA women offer to men, respectively (as calculated above).

```
men_high_sa <- partner_dist_het_sexact %>%
  filter(r_sex == "m" & r_sexact == "high")
format_table(men_high_sa)
```

r_sex	r_sexact	r_sexid	rp_sex	rp_sexact	r_demo	rp_demo	d_partners	prop
m	high	het	w	high	m_high	w_high	3.0419974	0.0722716
m	high	het	w	low	m_high	w_low	0.9648517	0.0722716

```
sum(men_high_sa$d_partners)
```

```
## [1] 4.006849
```

Notice that the partnerships are unbalanced. That is, if we calculate the total number of partnerships offered by each group, high SA men report a different number of partnerships with high SA women than high SA women do with high SA men.

```
partner_dist_het_sexact %>%
  mutate(total_group_pships = d_partners * prop) %>%
  filter(r_sexact == "high" & rp_sexact == "high") %>%
  select(r_demo, rp_demo, total_group_pships) %>%
  format_table()
```

r_demo	rp_demo	total_group_pships
m_high	w_high	0.2198501
w_high	m_high	0.0957605

Balancing

```r
natsal_het_bidi <- partner_dist_het_sexact %>%
    # do a self-join t calculate partners from perspective of rp
    inner_join(partner_dist_het_sexact,
               by = c("r_sex" = "rp_sex",
                      "r_sexact" = "rp_sexact",
                      "rp_sex" = "r_sex",
                      "rp_sexact" = "r_sexact",
                      "r_demo" = "rp_demo",
                      "rp_demo" = "r_demo"),
               suffix = c('.r', '.rp'))

theta <- 0.5
b_natsal_het_bidi <- natsal_het_bidi %>%
    mutate(np_r = d_partners.r * prop.r,
           np_rp = d_partners.rp * prop.rp) %>%
    mutate(imbalance = np_r / np_rp,
           corrected_r = d_partners.r / imbalance^(1 - theta),
           corrected_rp = d_partners.rp * imbalance^theta,
           cnr = corrected_r * prop.r,
           cnrp = corrected_rp * prop.rp) %>%
    select(r_demo,
           rp_demo,
           prop.r,
           prop.rp,
           corrected_r,
           corrected_rp,
           cnr,
           cnrp)

format_table(b_natsal_het_bidi)
```

r_demo	rp_demo	prop.r	prop.rp	corrected_r	corrected_rp	cnr	cnrp
m_high	w_high	0.0722716	0.0449038	2.0076530	3.2312674	0.1450963	0.1450963
m_low	w_high	0.4296301	0.0449038	0.1387567	1.3275934	0.0596141	0.0596141
m_high	w_low	0.0722716	0.4531944	0.8248611	0.1315419	0.0596141	0.0596141
m_low	w_low	0.4296301	0.4531944	0.0570094	0.0540451	0.0244929	0.0244929
w_high	m_high	0.0449038	0.0722716	3.2312674	2.0076530	0.1450963	0.1450963
w_low	m_high	0.4531944	0.0722716	0.1315419	0.8248611	0.0596141	0.0596141
w_high	m_low	0.0449038	0.4296301	1.3275934	0.1387567	0.0596141	0.0596141
w_low	m_low	0.4531944	0.4296301	0.0540451	0.0570094	0.0244929	0.0244929

Balanced! Let's check that this worked.

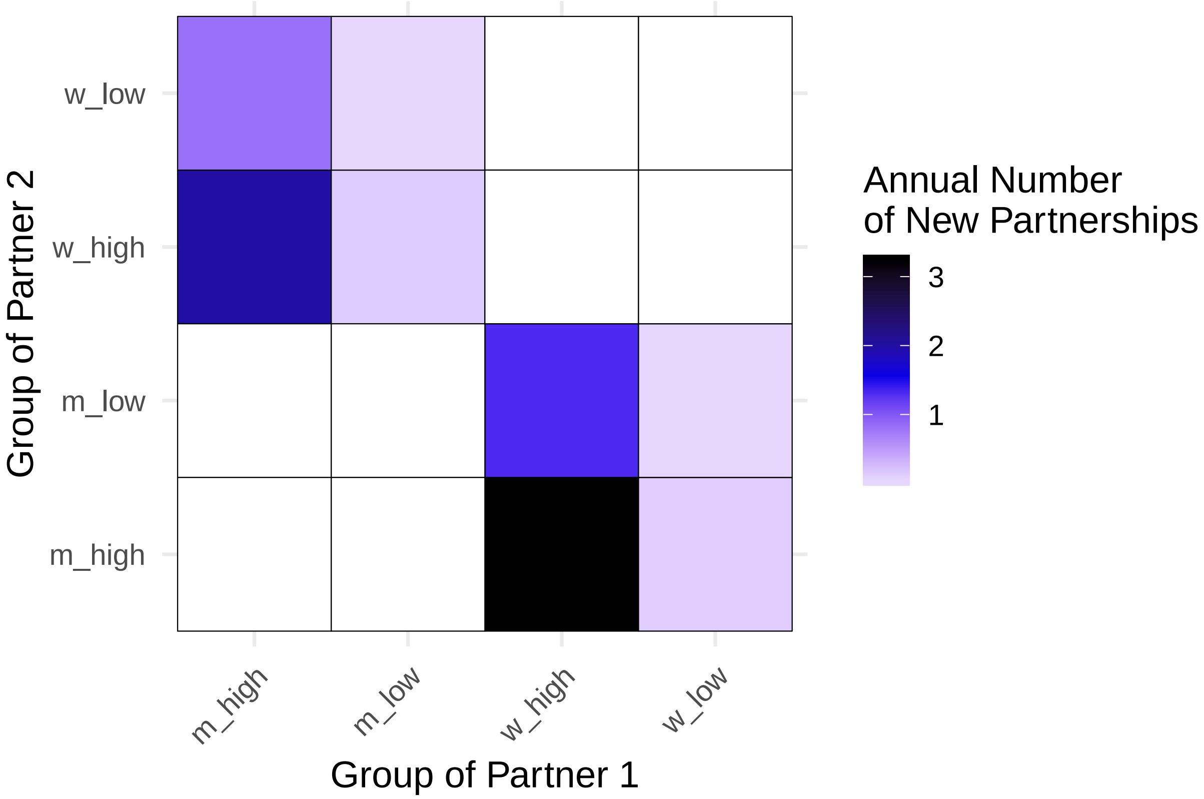
```r
all(with(b_natsal_het_bidi, abs(cnr - cnrp) < .Machine$double.eps))
```

```
## [1] TRUE
```

It works.

```r
# for plotting, I want to emphasize the zeros between same-sex
na_demos <- mutate(b_natsal_het_bidi,
                   r_demo = r_demo,
                   rp_demo = rev(rp_demo),
                    corrected_r = NA)

# plot with nas
library(ggplot2)
ggplot(rbind(b_natsal_het_bidi, na_demos)) +
    geom_tile(aes(x = r_demo, y = rp_demo, fill=corrected_r), color="black", size = 0.2) +
    scale_fill_gradient2(name="Annual Number\nof New Partnerships",
                         low = "white",
                         mid = "blue2",
                         high = "black",
                         midpoint = 1.5,
                         na.value = "white",
                         breaks = seq(0, 6)) +
    labs(x = "Group of Partner 1",
         y = "Group of Partner 2") +
    theme_minimal(base_size = 14) + coord_fixed() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Het, gay/lesbian, bi population, sexual activity stratification only

In this model, we do not stratify by sexual identity or restrict to heterosexual sexual identity. The analysis procedure is exactly the same as for the heterosexual-only population.

Estimate the number of new partnerships with each sex

```r
all_sexact_rep <- all_sex %>%
    group_by(r_sex, r_sexact, r_sexid) %>%
    summarise(partners = weighted.mean(hetnonew, w = total_wt)) %>%
    mutate(rp_sex = ifelse(r_sex == "m", 'w', 'm')) %>%
    rowwise() %>%
    mutate(prop = get_proportion(all_sex,
                                 q_sex = r_sex,
                                 q_sexact = r_sexact,
                                 q_sexid = r_sexid),
           n_partners = partners * prop) %>%
    ungroup()

format_table(all_sexact_rep)
```

r_sex	r_sexact	r_sexid	partners	rp_sex	prop	n_partners
m	high	bi	2.1793464	w	0.0012274	0.0026750
m	high	gay	0.2053960	w	0.0041569	0.0008538
m	high	het	4.0068490	w	0.0697660	0.2795418
m	low	bi	0.1108171	w	0.0036405	0.0004034
m	low	gay	0.0000000	w	0.0063202	0.0000000
m	low	het	0.2137853	w	0.4147349	0.0886642
w	high	bi	4.5772158	m	0.0024179	0.0110673
w	high	gay	0.5279848	m	0.0010434	0.0005509
w	high	het	3.2675397	m	0.0433470	0.1416382
w	low	bi	0.2242447	m	0.0101149	0.0022682
w	low	gay	0.0000000	m	0.0057486	0.0000000
w	low	het	0.1723069	m	0.4374822	0.0753812

Next, let's distribute the partnerships across activity levels.

Distribute partnerships

Since the respondent's don't know their partners' activity levels, we can make the proportionality assumption and then estimate the proportion of partnerships from each activity level.

This amounts to

$$\Pr(S_{rp} = s' | S_r = s) = \frac{N_{s'} \beta_{s's}}{\sum_i N_i \beta_{is}}$$

where S_c and S_p are the sex of the 'chooser' and 'partner' respectively, I_c and I_p are the the sexual identity of the 'chooser' and 'partner', N_{ks} is the proportion of people with sex k who have sexual identity s, and $\beta_{ksk'}$ is the number of partnerships that people of sex k with sexual identity s reported with sex k'.

We can do this using dplyr and piping. Basically, for each combination of `r_sex` and `rp_sex`, we take the total number of partnerships ‘offered’ by each sex and sexual activity group, and divide it by the total number of partnerships offered by that sex. Then, we define `rp_sexid = r_sexid`, so the proportion `prop_to_r` represents the proportion of partnerships from `rp_sex` offered to `r_sex` that come from people with sex `rp_sex` and SA group `rp_sexact`.

```
make_offered_dists_het_sexact <- function(df) {
  df %>%
    group_by(rp_sex, r_sex) %>%
    mutate(prop_to_r = n_partners / sum(n_partners)) %>%
    select(r_sex,
           rp_sex,
           rp_sexact = r_sexact,
           prop_to_r) %>%
    ungroup()
}
het_sexact_offered_dist <- make_offered_dists_het_sexact(het_sexact_rep)
format_table(het_sexact_offered_dist)
```

r_sex	rp_sex	rp_sexact	prop_to_r
m	w	high	0.7591994
m	w	low	0.2408006
w	m	high	0.6526522
w	m	low	0.3473478

As an example, the above data shows that women with high sexual actvity account for 75.9% of all partnerships “offered” to men, while women with low sexual activity account for 24.1%.

Now, we distribute the partnerships across the groups. To do this, we join the proportion dataframe with the survey dataframe. This join aligns the `prop_to_r` and `partners` column, so they can be multiplied.

```
joined_dfs <- left_join(het_sexact_offered_dist, het_sexact_rep, by = c("r_sex", "rp_sex"))
format_table(joined_dfs)
```

r_sex	rp_sex	rp_sexact	prop_to_r	r_sexact	r_sexid	partners	prop	n_partners
m	w	high	0.7591994	high	het	4.0068490	0.0722716	0.2895815
m	w	high	0.7591994	low	het	0.2137853	0.4296301	0.0918486
m	w	low	0.2408006	high	het	4.0068490	0.0722716	0.2895815
m	w	low	0.2408006	low	het	0.2137853	0.4296301	0.0918486
w	m	high	0.6526522	high	het	3.2675397	0.0449038	0.1467251
w	m	high	0.6526522	low	het	0.1723069	0.4531944	0.0780885
w	m	low	0.3473478	high	het	3.2675397	0.0449038	0.1467251
w	m	low	0.3473478	low	het	0.1723069	0.4531944	0.0780885

The multiplication gives us the distributed parterships. Then, we define a single variable that describes the two demographic variables (sex and SA group):

```
partner_dist_het_sexact <- joined_dfs %>%
  mutate(d_partners = partners * prop_to_r,
         r_demo = paste(r_sex, r_sexact, sep="_"),
         rp_demo = paste(rp_sex, rp_sexact, sep="_")) %>%
  select(r_sex, r_sexact, r_sexid, rp_sex, rp_sexact, r_demo, rp_demo, d_partners, prop)
format_table(partner_dist_het_sexact)
```

r_sex	r_sexact	r_sexid	rp_sex	rp_sexact	r_demo	rp_demo	d_partners	prop
m	high	het	w	high	m_high	w_high	3.0419974	0.0722716
m	low	het	w	high	m_low	w_high	0.1623057	0.4296301
m	high	het	w	low	m_high	w_low	0.9648517	0.0722716
m	low	het	w	low	m_low	w_low	0.0514796	0.4296301
w	high	het	m	high	w_high	m_high	2.1325671	0.0449038
w	low	het	m	high	w_low	m_high	0.1124565	0.4531944
w	high	het	m	low	w_high	m_low	1.1349726	0.0449038
w	low	het	m	low	w_low	m_low	0.0598504	0.4531944

Finally, we have to balance these partnerships.

```
natsal_het_bidi <- partner_dist_het_sexact %>%
    # do a self-join t calculate partners from perspective of rp
    inner_join(partner_dist_het_sexact,
               by = c("r_sex" = "rp_sex",
                      "r_sexact" = "rp_sexact",
                      "rp_sex" = "r_sex",
                      "rp_sexact" = "r_sexact",
                      "r_demo" = "rp_demo",
                      "rp_demo" = "r_demo"),
               suffix = c('.r', '.rp'))

theta <- 0.5
b_natsal_het_bidi <- natsal_het_bidi %>%
    mutate(np_r = d_partners.r * prop.r,
           np_rp = d_partners.rp * prop.rp) %>%
    mutate(imbalance = np_r / np_rp,
           corrected_r = d_partners.r / imbalance^(1 - theta),
           corrected_rp = d_partners.rp * imbalance^theta,
           cnr = corrected_r * prop.r,
           cnrp = corrected_rp * prop.rp) %>%
    select(r_demo,
           rp_demo,
           prop.r,
           prop.rp,
           corrected_r,
           corrected_rp,
           cnr,
           cnrp)

format_table(b_natsal_het_bidi)
```

r_demo	rp_demo	prop.r	prop.rp	corrected_r	corrected_rp	cnr	cnrp
m_high	w_high	0.0722716	0.0449038	2.0076530	3.2312674	0.1450963	0.1450963
m_low	w_high	0.4296301	0.0449038	0.1387567	1.3275934	0.0596141	0.0596141
m_high	w_low	0.0722716	0.4531944	0.8248611	0.1315419	0.0596141	0.0596141
m_low	w_low	0.4296301	0.4531944	0.0570094	0.0540451	0.0244929	0.0244929
w_high	m_high	0.0449038	0.0722716	3.2312674	2.0076530	0.1450963	0.1450963
w_low	m_high	0.4531944	0.0722716	0.1315419	0.8248611	0.0596141	0.0596141
w_high	m_low	0.0449038	0.4296301	1.3275934	0.1387567	0.0596141	0.0596141
w_low	m_low	0.4531944	0.4296301	0.0540451	0.0570094	0.0244929	0.0244929

Balanced! Let's check that this worked.

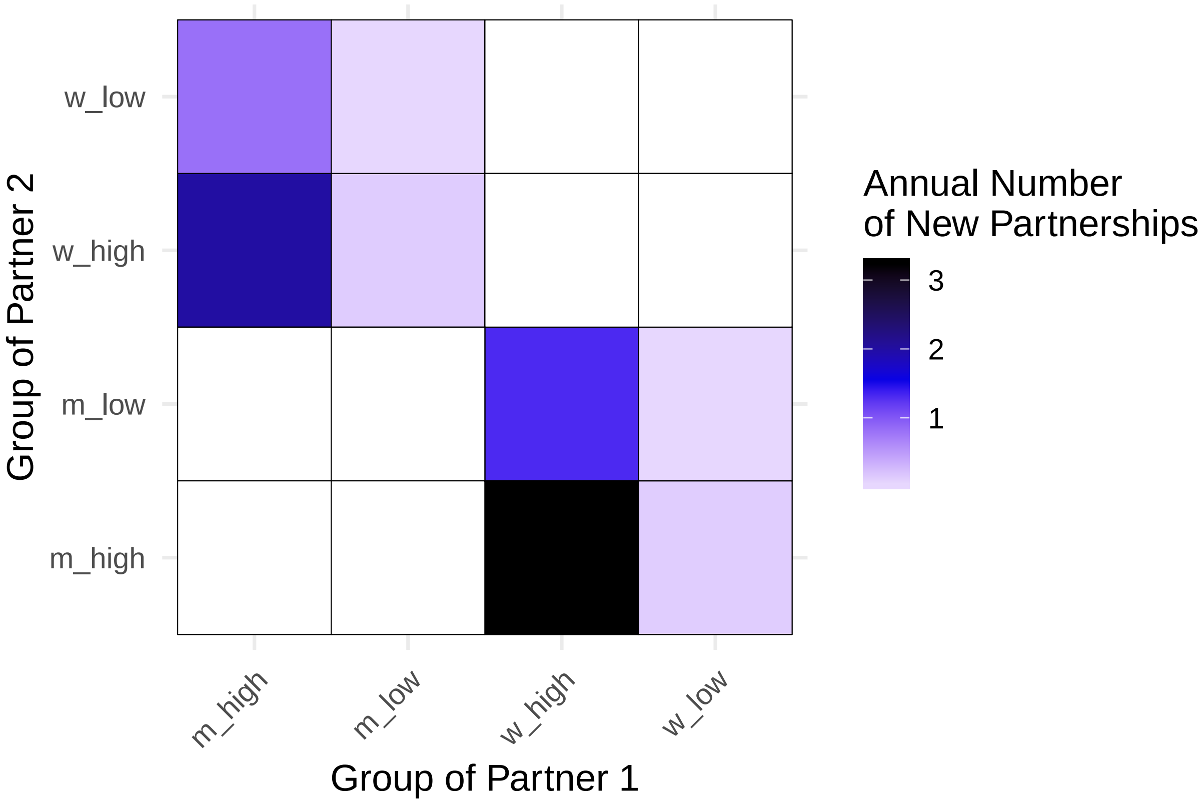
```
all(with(b_natsal_het_bidi, abs(cnr - cnrp) < .Machine$double.eps))
```

```
## [1] TRUE
```

It works.

```
# for plotting, I want to emphasize the zeros between same-sex
na_demos <- mutate(b_natsal_het_bidi,
                   r_demo = r_demo,
                   rp_demo = rev(rp_demo),
                   corrected_r = NA)

# plot with nas
library(ggplot2)
ggplot(rbind(b_natsal_het_bidi, na_demos)) +
    geom_tile(aes(x = r_demo, y = rp_demo, fill=corrected_r), color="black", size = 0.2) +
    scale_fill_gradient2(name="Annual Number\nof New Partnerships",
                         low = "white",
                         mid = "blue2",
                         high = "black",
                         midpoint = 1.5,
                         na.value = "white",
                         breaks = seq(0, 6)) +
    labs(x = "Group of Partner 1",
         y = "Group of Partner 2") +
    theme_minimal(base_size = 14) + coord_fixed() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```


w_low
w_high
m_low
m_high
Group of Partner 2
m_high
m_low
w_high
w_low
Group of Partner 1
Annual Number
of New Partnerships
3
2
1