Share of Environmental Exposure Burden Across Population Groups

Fan Wang

2021-12-27

Contents

1 Location, Population, and Pollution

Go to the **RMD**, **R**, **PDF**, or **HTML** version of this file. Go back to fan's REconTools Package, R Code Examples Repository (bookdown site), or Intro Stats with R Repository (bookdown site).

1.1 Location, Population, and Environmental Exposures (Pollution)

1.1.1 Local Public Exposures $Z_{l,y}$

Environmental exposure, specifically PM 2.5, is a local public good (bad). Prior research has shown that due to the nature of particular matter formation in the air, while there is particular matter variation at the city level, there is not significant variation in particular matter pollution at the neighborhood level He et al. (2019) and Liu et al. (2022). Additionally, there is a difference between ambient environmental exposures and the amount of particular matter inhaled by individual residents: the former is a common level shared by all residents, but the latter can differ depending on individuals' physical and socio-economic attributes.

There are M different locations (cities/counties/townships), indexed from $l = 1$ to $l = M$. In a particular time-period, we assume that the potential pollution exposure for all residents in the same location. Suppose additionally that we compute statistics at the interval of years y, and each year includes daily pollution measure for each day of the year t, indexed from $t = 1$ to $t = T_y$, where T_y is the total number of days in a year.

Let $\mathrm{Z}_{l,y}$ be the total pollution exposure by a resident in location l. This is equal to the sum of pollution exposure during the course of a year:

$$Z_{l,y} = \sum_{t=1}^{T_y} Z_{l,y,t}$$

1.1.2 Group Exposures $\mathcal{Z}_{i,y}$

Let there be N population groups, indexed from $i = 1$ to $i = N$. The N population groups reside in the M locations. the share, let $P_{l,i}$ denote share of population belong to group i that resides in location l:

$$1 = \sum_{l=1}^{M} \sum_{i=1}^{N} P_{l,i}$$

The share of population belong to population group i is:

$$P_i = \sum_{l=1}^{M} P_{l,i}$$

$\mathcal{Z}_{i,y}$, which is the average pollution exposure facing an individual belonging to group i in year y, is determined by how individuals from population group i are distributed across the M locations:

$$\mathcal{Z}_{i,y} = \sum_{l=1}^{M} \left(\frac{P_{l,i}}{P_i} \times Z_{l,y} \right)$$

Additionally, $\mathcal{Z}_y$ is the average pollution exposure facing an individual, regardless of population group, in year y:

$$\mathcal{Z}_y = \sum_{i=1}^{N} \left(\sum_{l=1}^{M} \left(\frac{P_{l,i}}{P_i} \times Z_{l,y} \right) P_i \right) = \sum_{i=1}^{N} \left(\sum_{l=1}^{M} \left(P_{l,i} \times Z_{l,y} \right) \right)$$

1.1.3 Excess Environmental Exposure Burden to Population

We have $\mathcal{Z}_{i,y}$, the average pollution burden facing an individual in a particular population group. We also have P_i, the share of population belong to population group i.

How does the share of pollution burden facing a population group relate to the share of population this group has in the overall population?

We define $\mathcal{E}_{i,y}$ as the share of pollution burden for population group i that is in excess of its population share as:

$$\mathcal{E}_{i,y} = \left(\frac{\mathcal{Z}_{i,y} \times P_i}{\sum_{\widehat{i}}^{N} \left(\mathcal{Z}_{\widehat{i},y} \times P_{\widehat{i}} \right)} \right) \times \frac{1}{P_i} = \frac{\mathcal{Z}_{i,y}}{\sum_{\widehat{i}}^{N} \left(\mathcal{Z}_{\widehat{i},y} \times P_{\widehat{i}} \right)}$$

1.2 Simulate Population Distribution over Location and Demographics

Use the binomial distribution to generate heterogenous demographic break-down by location. There are N demographic cells, and the binomial distribution provides the probability mass in each of the N cell. Different bernoulli "win" chance for each location. There is also probability distribution over population in each location.

First, construct empty population share dataframe:

```r
# 7 different age groups and 12 different locationso
it_N_pop_groups <- 7
it_M_location <- 12
# Matrix of demographics by location
mt_pop_data_frac <- matrix(data=NA, nrow=it_M_location, ncol=it_N_pop_groups)
colnames(mt_pop_data_frac) <- paste0('popgrp', seq(1,it_N_pop_groups))
rownames(mt_pop_data_frac) <- paste0('location', seq(1,it_M_location))
# Display
mt_pop_data_frac %>% kable() %>% kable_styling_fc()
```

	popgrp1	popgrp2	popgrp3	popgrp4	popgrp5	popgrp6	popgrp7
location1	NA	NA	NA	NA	NA	NA	NA
location2	NA	NA	NA	NA	NA	NA	NA
location3	NA	NA	NA	NA	NA	NA	NA
location4	NA	NA	NA	NA	NA	NA	NA
location5	NA	NA	NA	NA	NA	NA	NA
location6	NA	NA	NA	NA	NA	NA	NA
location7	NA	NA	NA	NA	NA	NA	NA
location8	NA	NA	NA	NA	NA	NA	NA
location9	NA	NA	NA	NA	NA	NA	NA
location10	NA	NA	NA	NA	NA	NA	NA
location11	NA	NA	NA	NA	NA	NA	NA
location12	NA	NA	NA	NA	NA	NA	NA

Second, generate conditional population distribution for each location, and then multiply by the share of population in each locality:

```r
# Share of population per location
set.seed(123)
ar_p_loc <- dbinom(0:(3*it_M_location-1), 3*it_M_location-1, 0.5)
it_start <- length(ar_p_loc)/2-it_M_location/2
ar_p_loc <- ar_p_loc[it_start:(it_start+it_M_location+1)]
ar_p_loc <- ar_p_loc/sum(ar_p_loc)

# Different bernoulli "win" probability for each location
set.seed(234)
# ar_fl_unif_prob <- sort(runif(it_M_location)*(0.25)+0.4)
ar_fl_unif_prob <- sort(runif(it_M_location))

# Generate population proportion by locality
for (it_loc in 1:it_M_location ) {
  ar_p_pop_condi_loc <- dbinom(0:(it_N_pop_groups-1), it_N_pop_groups-1, ar_fl_unif_prob[it_loc])
  mt_pop_data_frac[it_loc,] <- ar_p_pop_condi_loc*ar_p_loc[it_loc]
}

# Sum of cells, should equal to 1
print(paste0('pop frac sum = ', sum(mt_pop_data_frac)))
```

```
## [1] "pop frac sum = 0.962953679726938"
```

```r
# Display
round(mt_pop_data_frac*100, 2) %>%
  kable(caption='Share of population in each location and demographic cell') %>%
  kable_styling_fc()
```

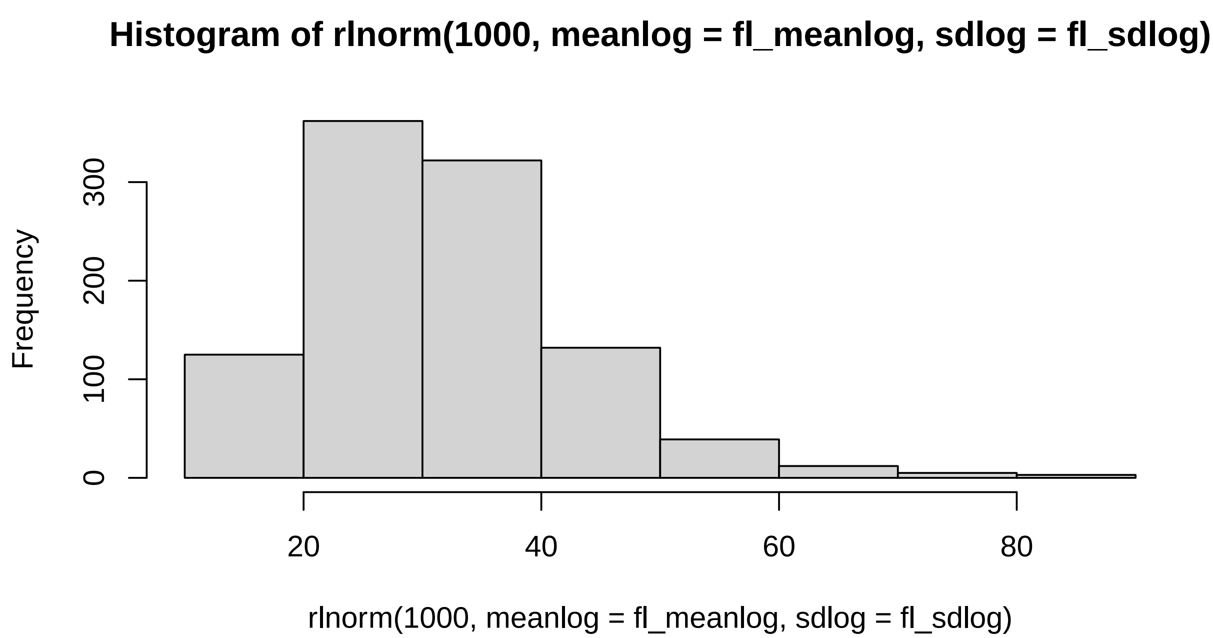
Share of population in each location and demographic cell

	popgrp1	popgrp2	popgrp3	popgrp4	popgrp5	popgrp6	popgrp7
location1	1.09	0.13	0.01	0.00	0.00	0.00	0.00
location2	1.63	0.70	0.13	0.01	0.00	0.00	0.00
location3	0.59	1.40	1.39	0.74	0.22	0.03	0.00
location4	0.06	0.43	1.29	2.09	1.90	0.92	0.19
location5	0.07	0.55	1.73	2.89	2.71	1.36	0.28
location6	0.02	0.26	1.19	2.89	3.93	2.85	0.86
location7	0.01	0.10	0.66	2.23	4.26	4.33	1.83
location8	0.00	0.06	0.47	1.83	4.03	4.72	2.31
location9	0.00	0.03	0.27	1.26	3.28	4.55	2.63
location10	0.00	0.02	0.20	0.96	2.57	3.68	2.19
location11	0.00	0.00	0.00	0.04	0.40	2.05	4.38
location12	0.00	0.00	0.00	0.02	0.24	1.28	2.82

1.3 Simulate Environmental Exposure

Use log-normal distribution to describe average daily PM10 exposures distribution by locality:

```
fl_meanlog <- 3.4
fl_sdlog <- 0.35
hist(rlnorm(1000, meanlog = fl_meanlog, sdlog = fl_sdlog))
```



First, draw pollution measure for each locality:

```
# draw
set.seed(123)
ar_pollution_loc <- rlnorm(it_M_location, meanlog = fl_meanlog, sdlog = fl_sdlog)
# pollution dataframe
# 5 by 3 matrix
```

```
# Column Names
ar_st_varnames <- c('location','avgdailypm10')

# Combine to tibble, add name col1, col2, etc.
tb_loc_pollution <- as_tibble(ar_pollution_loc) %>%
  rowid_to_column(var = "id") %>%
  rename_all(~c(ar_st_varnames)) %>%
  mutate(location = paste0('location', location))

# Display
kable(tb_loc_pollution) %>% kable_styling_fc()
```

location	avgdailypm10
location1	24.62676
location2	27.64481
location3	51.70466
location4	30.71275
location5	31.35114
location6	54.61304
location7	35.20967
location8	19.24456
location9	23.56121
location10	25.63653
location11	45.99021
location12	33.98553

Second, reshape population data:

```
# Reshape population data, so each observation is location/demo
df_pop_data_frac_long <- as_tibble(mt_pop_data_frac, rownames='location') %>%
  pivot_longer(cols = starts_with('popgrp'),
               names_to = c('popgrp'),
               names_pattern = paste0("popgrp(.*)"),
               values_to = "pop_frac")
```

Third, join with pollution data:

```
# Reshape population data, so each observation is location/demo
df_pop_pollution_long <- df_pop_data_frac_long %>%
  left_join(tb_loc_pollution, by='location')

# display
df_pop_pollution_long[1:round(it_N_pop_groups*2.5),] %>% kable() %>% kable_styling_fc()
```

1.4 Compute Demographic Group Specific Exposure Distributions

What is the p10, median, p90 and mean pollution exposure for each demographic group?

1. group by population group
2. sort by pollution exposure within group
3. generate population group specific conditional population weights
4. generate population CDF for each population group (sorted by pollution)

location	popgrp	pop_frac	avgdailypm10
location1	1	0.0109366	24.62676
location1	2	0.0013417	24.62676
location1	3	0.0000686	24.62676
location1	4	0.0000019	24.62676
location1	5	0.0000000	24.62676
location1	6	0.0000000	24.62676
location1	7	0.0000000	24.62676
location2	1	0.0163003	27.64481
location2	2	0.0070132	27.64481
location2	3	0.0012573	27.64481
location2	4	0.0001202	27.64481
location2	5	0.0000065	27.64481
location2	6	0.0000002	27.64481
location2	7	0.0000000	27.64481
location3	1	0.0058760	51.70466
location3	2	0.0140000	51.70466
location3	3	0.0138984	51.70466
location3	4	0.0073587	51.70466

```
# Follow four steps above
df_pop_pollution_by_popgrp_cdf <- df_pop_pollution_long %>%
  arrange(popgrp, avgdailypm10) %>%
  group_by(popgrp) %>%
  mutate(cdf_pop_condi_popgrp_sortpm10 = cumsum(pop_frac/sum(pop_frac)),
         pmf_pop_condi_popgrp_sortpm10 = (pop_frac/sum(pop_frac)))
# display
df_pop_pollution_by_popgrp_cdf[1:round(it_N_pop_groups*5.5),] %>%
  kable() %>% kable_styling_fc_wide()
```

1.5 Various Relative Burden Statistics

What to compute?

1. Excess pollution burden: Share of pollution burden by population group and overall population share, this is simply the ratio of population group mean and the overall weighted mean.
2. What is the fraction of the people in each population group with below and above overall average?
3. Merge results for different quantiles together.

1.5.1 Excess pollution burden

We compute excess population burden: *pm10_grp_exc_burden*. 0.10 means 10 percent in excess, this means the pollution burden share is 10 percent in excess of the population share. -0.10 means 10 percent less than what population share is.

Additionally, we compute the share of people within group above the overall mean: *pm10_grp_shr_exc*. This shows the share of people having excess burden. This complements the first number. Because the 10 percent excess could be due to very high exposure to a very small number of people within a population group, or it could be that most people in the group are in "excess".

```
# Stats 1: excess pollution burden
df_excess_pollution_burden <- df_pop_pollution_by_popgrp_cdf %>%
  ungroup() %>%
  mutate(pm10_overall_mean = weighted.mean(avgdailypm10, pop_frac)) %>%
```

location	popgrp	pop_frac	avgdailypm10	cdf_pop_condi_popgrp_sortpm10	pmf_pop_condi_popgrp_sortpm10
location8	1	0.0000364	19.24456	0.0010453	0.0010453
location9	1	0.0000151	23.56121	0.0014804	0.0004351
location1	1	0.0109366	24.62676	0.3156484	0.3141680
location10	1	0.0000104	25.63653	0.3159471	0.0002988
location2	1	0.0163003	27.64481	0.7841942	0.4682471
location4	1	0.0005879	30.71275	0.8010816	0.0168874
location5	1	0.0007392	31.35114	0.8223166	0.0212350
location12	1	0.0000000	33.98553	0.8223168	0.0000002
location7	1	0.0000681	35.20967	0.8242718	0.0019550
location11	1	0.0000000	45.99021	0.8242721	0.0000003
location3	1	0.0058760	51.70466	0.9930669	0.1687948
location6	1	0.0002413	54.61304	1.0000000	0.0069331
location8	2	0.0006400	19.24456	0.0172871	0.0172871
location9	2	0.0003150	23.56121	0.0257947	0.0085076
location1	2	0.0013417	24.62676	0.0620374	0.0362427
location10	2	0.0002235	25.63653	0.0680736	0.0060362
location2	2	0.0070132	27.64481	0.2575157	0.1894421
location4	2	0.0042712	30.71275	0.3728918	0.1153760
location5	2	0.0055479	31.35114	0.5227547	0.1498629
location12	2	0.0000004	33.98553	0.5227662	0.0000116
location7	2	0.0010378	35.20967	0.5508009	0.0280347
location11	2	0.0000008	45.99021	0.5508213	0.0000203
location3	2	0.0140000	51.70466	0.9289930	0.3781718
location6	2	0.0026287	54.61304	1.0000000	0.0710070
location8	3	0.0046896	19.24456	0.0638166	0.0638166
location9	3	0.0027290	23.56121	0.1009539	0.0371373
location1	3	0.0000686	24.62676	0.1018872	0.0009333
location10	3	0.0020006	25.63653	0.1291118	0.0272246
location2	3	0.0012573	27.64481	0.1462207	0.0171089
location4	3	0.0129304	30.71275	0.3221799	0.1759592
location5	3	0.0173492	31.35114	0.5582709	0.2360910
location12	3	0.0000141	33.98553	0.5584625	0.0001916
location7	3	0.0065945	35.20967	0.6482016	0.0897391
location11	3	0.0000242	45.99021	0.6485305	0.0003290
location3	3	0.0138984	51.70466	0.8376617	0.1891312
location6	3	0.0119295	54.61304	1.0000000	0.1623383
location8	4	0.0183277	19.24456	0.1224562	0.1224562
location9	4	0.0126118	23.56121	0.2067219	0.0842656

```r
  group_by(popgrp) %>%
  mutate(pm10_grp_mean = weighted.mean(avgdailypm10, pop_frac)) %>%
  slice(1) %>%
  mutate(pm10_grp_exc_burden = pm10_grp_mean/pm10_overall_mean - 1) %>%
  select(popgrp, pm10_grp_mean, pm10_overall_mean, pm10_grp_exc_burden)
fl_pm10_overall_mean <- mean(df_excess_pollution_burden %>% pull(pm10_overall_mean))

# Stats 2: share of people within group below or above overall mean
df_share_below_or_excess <- df_pop_pollution_by_popgrp_cdf %>%
  arrange(popgrp, avgdailypm10) %>%
  filter(avgdailypm10 < fl_pm10_overall_mean) %>%
  slice_tail() %>%
  mutate(pm10_grp_shr_exc = 1 - cdf_pop_condi_popgrp_sortpm10) %>%
  select(popgrp, pm10_grp_shr_exc)
# merge stats 2 with stats 1
df_excess_pollution_burden <- df_excess_pollution_burden %>%
  left_join(df_share_below_or_excess, by="popgrp")
```

```r
# display
df_excess_pollution_burden %>%
  kable(caption = 'PM10 Exposure Distribution by Population Groups') %>%
  kable_styling_fc()
```

PM10 Exposure Distribution by Population Groups

popgrp	pm10_grp_mean	pm10_overall_mean	pm10_grp_exc_burden	pm10_grp_shr_exc
1	31.07894	33.42117	-0.0700825	0.1776834
2	39.47897	33.42117	0.1812562	0.4772453
3	37.92901	33.42117	0.1348797	0.4417291
4	34.86470	33.42117	0.0431920	0.3958166
5	32.56731	33.42117	-0.0255485	0.3846056
6	31.46626	33.42117	-0.0584934	0.4093719
7	33.50541	33.42117	0.0025205	0.5657246

1.5.2 Within Group Percentiles

Compute within group percentiles for each population groups. Use the list of percentiles below to specify which percentiles should be computed.

```r
# Stats 3: percentiles and ratios
ar_fl_percentiles <- c(0.1, 0.2, 0.8, 0.9)
# Stats 3a: generate key within group percentiles
# 1. 20th and 80th percentiles
# 2. 10th and 90th percentiles
# 3. 50th percentile
# Generate pollution quantiles by population groups
for (it_percentile_ctr in seq(1, length(ar_fl_percentiles))) {

    # Current within group percentile to compute
    fl_percentile <- ar_fl_percentiles[it_percentile_ctr]
    svr_percentile <- paste0('pm10_p', round(fl_percentile*100))

    # Frame with specific percentile
    df_within_percentiles_cur <- df_pop_pollution_by_popgrp_cdf %>%
      group_by(popgrp) %>%
      filter(cdf_pop_condi_popgrp_sortpm10 >= fl_percentile) %>%
      slice(1) %>%
      mutate(!!sym(svr_percentile) := avgdailypm10) %>%
      select(popgrp, one_of(svr_percentile))

    # Merge percentile frames together
    if (it_percentile_ctr > 1) {
      df_within_percentiles <- df_within_percentiles %>%
        left_join(df_within_percentiles_cur, by='popgrp')
    } else {
      df_within_percentiles <- df_within_percentiles_cur
    }
}

# display
df_within_percentiles %>%
  kable(caption = 'PM10 Exposure Distribution by Population Groups') %>%
```

```
  kable_styling_fc()
```

PM10 Exposure Distribution by Population Groups

popgrp	pm10_p10	pm10_p20	pm10_p80	pm10_p90
1	24.62676	24.62676	30.71275	51.70466
2	27.64481	27.64481	51.70466	51.70466
3	23.56121	30.71275	51.70466	54.61304
4	19.24456	23.56121	51.70466	54.61304
5	19.24456	23.56121	35.20967	54.61304
6	19.24456	23.56121	35.20967	54.61304
7	19.24456	23.56121	45.99021	45.99021

1.5.3 Percentiles As Excess Burden

The 80th percentile for a population group, how is this exposed relative to the mean? We simply divide the within group pollution percentiles by the overall mean across all groups. All are properly weighted.

This is relating within group percentiles to the overall mean. These can be interpreted as excess burdens at specific percentiles. Individuals at the 80th percentile of a particular population group, how does their pollution burden compare to their population share?

```
# merge stats 3 with stats 1 and 2
df_excess_pollution_burden <- df_excess_pollution_burden %>%
  left_join(df_within_percentiles, by="popgrp")

# Stats 3b: Percentiles to Relative Burdens
# Convert percentiles to be relative of overall means
for (it_percentile_ctr in seq(1, length(ar_fl_percentiles))) {

    # Current within group percentile to compute
    fl_percentile <- ar_fl_percentiles[it_percentile_ctr]
    svr_percentile <- paste0('pm10_p', round(fl_percentile*100))
    svr_perc_exc_burden <- paste0('pm10_grp_excbrd_p', round(fl_percentile*100))

    # Percentiles to excess percentiles
    df_excess_pollution_burden  <- df_excess_pollution_burden  %>%
      mutate(!!sym(svr_perc_exc_burden) := !!sym(svr_percentile)/pm10_overall_mean)
}

# display
df_excess_pollution_burden %>%
  select(-pm10_overall_mean,
         -starts_with('pm10_p')) %>%
  kable(caption = 'PM10 Exposure Distribution by Population Groups') %>%
  kable_styling_fc_wide()
```

PM10 Exposure Distribution by Population Groups

popgrp	pm10_grp_mean	pm10_grp_exc_burden	pm10_grp_shr_exc	pm10_grp_excbrd_p10	pm10_grp_excbrd_p20	pm10_grp_excbrd_p80	pm10_grp_excbrd_p90
1	31.07894	-0.0700825	0.1776834	0.7368609	0.7368609	0.9189609	1.547063
2	39.47897	0.1812562	0.4772453	0.8271644	0.8271644	1.5470629	1.547063
3	37.92901	0.1348797	0.4417291	0.7049784	0.9189609	1.5470629	1.634085
4	34.86470	0.0431920	0.3958166	0.5758195	0.7049784	1.5470629	1.634085
5	32.56731	-0.0255485	0.3846056	0.5758195	0.7049784	1.0535138	1.634085
6	31.46626	-0.0584934	0.4093719	0.5758195	0.7049784	1.0535138	1.634085
7	33.50541	0.0025205	0.5657246	0.5758195	0.7049784	1.3760801	1.376080

1.5.4 Within Group Relative Exposure Ratios Across Percentiles

We now compute within group relative ratios of interest. This is purely within group inequality.

```r
# lower and upper bound or relative within group ratios
# can only use values appearing in the percentiles list prior
ar_fl_ratio_upper <- c(0.8, 0.9)
ar_fl_ratio_lower <- c(0.2, 0.1)
# Stats 4c: Ratios
# Generate P80 to P20 ratio, and P90 to P10 standard inequality ratios
for (it_ratio_ctr in seq(1, length(ar_fl_ratio_upper))) {

    # Upper and lower percentile bounds
    fl_ratio_upper <- ar_fl_ratio_upper[it_ratio_ctr]
    fl_ratio_lower <- ar_fl_ratio_lower[it_ratio_ctr]
    svr_ratio_upper_perc <- paste0('pm10_p', round(fl_ratio_upper*100))
    svr_ratio_lower_perc <- paste0('pm10_p', round(fl_ratio_lower*100))

    # New relative within group ratio variable name
    svr_ratio <- paste0('pm10_rat_p', round(fl_ratio_upper*100), '_dvd_p', round(fl_ratio_lower*100))

    # Generate P80 to P20 ratio, etc.
    df_excess_pollution_burden  <- df_excess_pollution_burden  %>%
      mutate(!!sym(svr_ratio) := !!sym(svr_ratio_upper_perc)/!!sym(svr_ratio_lower_perc))
}

# display
df_excess_pollution_burden %>%
  select(-pm10_overall_mean,
         -starts_with('pm10_grp_excbrd_p'),
         -starts_with('pm10_p')) %>%
  kable(caption = 'PM10 Exposure Distribution by Population Groups') %>%
  kable_styling_fc_wide()
```

PM10 Exposure Distribution by Population Groups

popgrp	pm10_grp_mean	pm10_grp_exc_burden	pm10_grp_shr_exc	pm10_rat_p80_dvd_p20	pm10_rat_p90_dvd_p10
1	31.07894	-0.0700825	0.1776834	1.247129	2.099532
2	39.47897	0.1812562	0.4772453	1.870321	1.870321
3	37.92901	0.1348797	0.4417291	1.683492	2.317922
4	34.86470	0.0431920	0.3958166	2.194483	2.837843
5	32.56731	-0.0255485	0.3846056	1.494392	2.837843
6	31.46626	-0.0584934	0.4093719	1.494392	2.837843
7	33.50541	0.0025205	0.5657246	1.951947	2.389777