

# Summarizing Data Tutorial

Grace Arkfeld

2024-03-05

```
rm(list = ls())
```

```
# Load the tidyverse package  
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --  
## v dplyr      1.1.4      v readr      2.1.5  
## v forcats    1.0.0      v stringr   1.5.1  
## v ggplot2    3.4.4      v tibble    3.2.1  
## v lubridate  1.9.3      v tidyr     1.3.1  
## v purrr      1.0.2  
## -- Conflicts ----- tidyverse_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag()     masks stats::lag()  
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
# Load the knitr package  
library(knitr)
```

```
la_mort <-  
  read_csv("https://www.dropbox.com/scl/fi/fzsnhfd3lq80v2o3sag6c/la\_mort.csv?rlkey=h1vyjm2b8ppgejgsg3e8")
```

```
## Rows: 642696 Columns: 29  
## -- Column specification -----  
## Delimiter: ","  
## chr (7): stocr, strsd, stbrth, brthr, sex, marstat, ucod  
## dbl (22): restatus, cntyocr, popcntyocr, cntyrtd, popcntyresd, educ1989, edu...  
##  
## i Use 'spec()' to retrieve the full column specification for this data.  
## i Specify the column types or set 'show_col_types = FALSE' to quiet this message.
```

```
la_pop <-  
  read_csv("https://www.dropbox.com/scl/fi/650k1obpczky6bwa19ex6/la\_county\_pop.csv?rlkey=0aokd9m76q7mxw")
```

```
## Rows: 24320 Columns: 23  
## -- Column specification -----  
## Delimiter: ","  
## chr (3): stname, ctynome, agegrp  
## dbl (20): state, county, year, tot_pop, tot_male, tot_female, wa_male, wa_fe...  
##  
## i Use 'spec()' to retrieve the full column specification for this data.  
## i Specify the column types or set 'show_col_types = FALSE' to quiet this message.
```

```

stnrd_pop <-
  read_csv("https://www.dropbox.com/scl/fi/xzd2o5lza237so6vamqwb/stnrd_pop.csv?rlkey=zp90au2tuq6eptvi1y

## Rows: 18 Columns: 2
## -- Column specification -----
## Delimiter: ","
## chr (1): agegrp
## dbl (1): stnrd_pop
##
## i Use 'spec()' to retrieve the full column specification for this data.
## i Specify the column types or set 'show_col_types = FALSE' to quiet this message.

# Step 4: Define Cancer Alley Parishes
la_mort$cancer_alley_parish <- ifelse(la_mort$cntyrstd %in% c(5, 47, 89, 93, 95, 121), 1, 0)

# Step 5: Define Cancer Deaths by Cancer Site
la_mort$stomach_cancer <- ifelse(la_mort$ucr39 == 5, 1, 0)
la_mort$colon_cancer <- ifelse(la_mort$ucr39 == 6, 1, 0)
la_mort$pancreas_cancer <- ifelse(la_mort$ucr39 == 7, 1, 0)
la_mort$lung_cancer <- ifelse(la_mort$ucr39 == 8, 1, 0)
la_mort$breast_cancer <- ifelse(la_mort$ucr39 == 9, 1, 0)
la_mort$cervix_cancer <- ifelse(la_mort$ucr39 == 10, 1, 0)
la_mort$prostate_cancer <- ifelse(la_mort$ucr39 == 11, 1, 0)
la_mort$bladder_cancer <- ifelse(la_mort$ucr39 == 12, 1, 0)
la_mort$lymphoma_cancer <- ifelse(la_mort$ucr39 == 13, 1, 0)
la_mort$leukemia_cancer <- ifelse(la_mort$ucr39 == 14, 1, 0)
la_mort$other_site_cancer <- ifelse(la_mort$ucr39 == 15, 1, 0)
la_mort$total_cancer <- ifelse(la_mort$ucr39 %in% 5:15, 1, 0)

# Step 6: Adjust Age Groupings
la_mort <- la_mort %>% filter(age != 9999)
age_breaks <- c(0, 1, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, 105, 110, 115, 120, 125, 130, 135, 140, 145, 150)
age_labels <- c("0", "1-4", "5-9", "10-14", "15-19", "20-24", "25-29", "30-34", "35-39", "40-44", "45-49", "50-54", "55-59", "60-64", "65-69", "70-74", "75-79", "80-84", "85-89", "90-94", "95-99", "100-104", "105-109", "110-114", "115-119", "120-124", "125-129", "130-134", "135-139", "140-144", "145-149", "150-154", "155-159", "160-164", "165-169", "170-174", "175-179", "180-184", "185-189", "190-194", "195-199", "200-204", "205-209", "210-214", "215-219", "220-224", "225-229", "230-234", "235-239", "240-244", "245-249", "250-254", "255-259", "260-264", "265-269", "270-274", "275-279", "280-284", "285-289", "290-294", "295-299", "300-304", "305-309", "310-314", "315-319", "320-324", "325-329", "330-334", "335-339", "340-344", "345-349", "350-354", "355-359", "360-364", "365-369", "370-374", "375-379", "380-384", "385-389", "390-394", "395-399", "400-404", "405-409", "410-414", "415-419", "420-424", "425-429", "430-434", "435-439", "440-444", "445-449", "450-454", "455-459", "460-464", "465-469", "470-474", "475-479", "480-484", "485-489", "490-494", "495-499", "500-504", "505-509", "510-514", "515-519", "520-524", "525-529", "530-534", "535-539", "540-544", "545-549", "550-554", "555-559", "560-564", "565-569", "570-574", "575-579", "580-584", "585-589", "590-594", "595-599", "600-604", "605-609", "610-614", "615-619", "620-624", "625-629", "630-634", "635-639", "640-644", "645-649", "650-654", "655-659", "660-664", "665-669", "670-674", "675-679", "680-684", "685-689", "690-694", "695-699", "700-704", "705-709", "710-714", "715-719", "720-724", "725-729", "730-734", "735-739", "740-744", "745-749", "750-754", "755-759", "760-764", "765-769", "770-774", "775-779", "780-784", "785-789", "790-794", "795-799", "800-804", "805-809", "810-814", "815-819", "820-824", "825-829", "830-834", "835-839", "840-844", "845-849", "850-854", "855-859", "860-864", "865-869", "870-874", "875-879", "880-884", "885-889", "890-894", "895-899", "900-904", "905-909", "910-914", "915-919", "920-924", "925-929", "930-934", "935-939", "940-944", "945-949", "950-954", "955-959", "960-964", "965-969", "970-974", "975-979", "980-984", "985-989", "990-994", "995-999")
la_mort$agegrp <- cut(la_mort$age, breaks = age_breaks, labels = age_labels, right = FALSE)

# Step 7: Define Race in the Mortality File
la_mort <- la_mort %>% filter(racer5 %in% c(1, 2))
la_mort$black <- ifelse(la_mort$racer5 == 2, 1, 0)

# Step 8: Create Parish Counts of Cancer Deaths by Cancer Site and by Race
parish_counts <- la_mort %>%
  group_by(cntyrstd, black, year, agegrp) %>%
  summarize(
    stomach_cancer_deaths = sum(stomach_cancer, na.rm = TRUE),
    colon_cancer_deaths = sum(colon_cancer, na.rm = TRUE),
    pancreas_cancer_deaths = sum(pancreas_cancer, na.rm = TRUE),
    lung_cancer_deaths = sum(lung_cancer, na.rm = TRUE),
    breast_cancer_deaths = sum(breast_cancer, na.rm = TRUE),
    cervix_cancer_deaths = sum(cervix_cancer, na.rm = TRUE),
    prostate_cancer_deaths = sum(prostate_cancer, na.rm = TRUE),
    bladder_cancer_deaths = sum(bladder_cancer, na.rm = TRUE),
    lymphoma_cancer_deaths = sum(lymphoma_cancer, na.rm = TRUE),

```

```

    leukemia_cancer_deaths = sum(leukemia_cancer, na.rm = TRUE),
    other_site_cancer_deaths = sum(other_site_cancer, na.rm = TRUE),
    total_cancer_deaths = sum(total_cancer, na.rm = TRUE)
  )

```

## 'summarise()' has grouped output by 'cntyrstd', 'black', 'year'. You can  
## override using the '.groups' argument.

*# Step 9: Define Race in the Population File*

```

la_pop <- la_pop %>%
  mutate(
    black_pop = rowSums(select(., c("ba_male", "ba_female"))),
    white_pop = rowSums(select(., c("wa_male", "wa_female"))),
  )
la_pop_black <- select(la_pop, county, year, agegrp, black_pop)
la_pop_white <- select(la_pop, county, year, agegrp, white_pop)

```

*# Step 10: Join the Mortality and Population Data Frames*

```

la_joined_black <- parish_counts %>%
  filter(black == 1) %>%
  inner_join(la_pop_black, by = c("cntyrstd" = "county", "year", "agegrp")) %>%
  rename(tot_pop = black_pop)

la_joined_white <- parish_counts %>%
  filter(black == 0) %>%
  inner_join(la_pop_white, by = c("cntyrstd" = "county", "year", "agegrp")) %>%
  rename(tot_pop = white_pop)

la_bind <- rbind(la_joined_black, la_joined_white)

```

*# Step 11: Join the Mortality/Population Data to the Standard Population Data*

*# Assuming stnrd\_pop is the standard population data frame*

```

la_bind <- la_bind %>%
  inner_join(stnrd_pop, by = "agegrp")

```

*# Step 12: Calculate Population Weights*

```

la_bind$stnrd_pop_weight <- (la_bind$stnrd_pop) / (sum(stnrd_pop$stnrd_pop))

```

*# Step 13: Calculate Cancer Mortality Rates by Cancer Site and Race*

```

cancer_sites <- c("stomach", "colon", "pancreas", "lung", "breast", "cervix", "prostate", "bladder", "liver")
for (site in cancer_sites) {
  rate_col <- paste(site, "cancer_rate_adj", sep = "_")
  death_col <- paste(site, "cancer_deaths", sep = "_")
  la_bind[[rate_col]] <- ((la_bind[[death_col]]) / (la_bind$tot_pop / 100000)) * la_bind$stnrd_pop_weight
}

```

*# Replace "inf" values with NA*

```

for (col in names(la_bind)) {
  la_bind[[col]][is.infinite(la_bind[[col]])] <- NA
}

```

```
# Step 14: Aggregate to the Parish-Year Level
parish_rates <- la_bind %>%
  group_by(cntyrds, black, year) %>%
  summarize(across(ends_with("cancer_rate_adj"), sum, na.rm = TRUE), tot_pop = sum(tot_pop))
```

```
## Warning: There was 1 warning in 'summarize()'.
## i In argument: 'across(ends_with("cancer_rate_adj"), sum, na.rm = TRUE)'.
## Caused by warning:
## ! The '...' argument of 'across()' is deprecated as of dplyr 1.1.0.
## Supply arguments directly to '.fns' through an anonymous function instead.
##
## # Previously
##   across(a:b, mean, na.rm = TRUE)
##
## # Now
##   across(a:b, \(x) mean(x, na.rm = TRUE))

## 'summarise()' has grouped output by 'cntyrds', 'black'. You can override using
## the '.groups' argument.
```

```
# Step 15: Weight by Parish Population
for (site in cancer_sites) {
  rate_col <- paste(site, "cancer_rate_adj", sep = "_")
  weight_col <- paste(site, "pop_weight", sep = "_")
  parish_rates[[weight_col]] <- parish_rates[[rate_col]] * parish_rates$tot_pop
}
```

```
# Step 15: Weight by Parish Population
for (site in cancer_sites) {
  rate_col <- paste(site, "cancer_rate_adj", sep = "_")
  weight_col <- paste(site, "pop_weight", sep = "_")
  parish_rates[[weight_col]] <- parish_rates[[rate_col]] * parish_rates$tot_pop
}
```

```
# Create a data frame with cancer_alley_parish information
cancer_alley_info <- la_mort %>%
  select(cntyrds, cancer_alley_parish) %>%
  distinct()
```

```
# Add cancer_alley_parish information to parish_rates
parish_rates <- parish_rates %>%
  left_join(cancer_alley_info, by = "cntyrds")
```

```
# Step 16: Aggregate to Cancer Alley and non-Cancer Alley Parishes
cancer_alley_rates <- parish_rates %>%
  group_by(cancer_alley_parish, black, year) %>%
  summarize(across(ends_with("pop_weight"), sum), tot_pop = sum(tot_pop)) %>%
  mutate(across(ends_with("pop_weight"), ~ .x / tot_pop))
```

```
## 'summarise()' has grouped output by 'cancer_alley_parish', 'black'. You can
## override using the '.groups' argument.
```