urban_rural.rmd

Madison Thompson

2026-02-03

```
# load libraries
library(dplyr)
library(readr)
library(tidyr)

nc_health_data_final <- read_csv("nc_county_health_data_final.csv")
nc_health_data_2021 <- read_csv("nc_county_health_data_2021_cleaned.csv")

# 2021 data is transposed (counties as columns). Extract just the Metro row,
# then pivot so counties become rows and we can merge on County.
metro_row <- nc_health_data_2021 %>%
  filter(Metric == "Metropolitan or Nonmetropolitan") %>%
  select(-Metric, -NC) %>%
  pivot_longer(cols = everything(),
               names_to = "County",
               values_to = "Metro_Status") %>%
  mutate(Metro_Status = trimws(Metro_Status),
         Metro_Status = na_if(Metro_Status, ""),
        Metro_Status = gsub("Nometro", "Nonmetro", Metro_Status))

# Merge onto final data
nc_health_data_merged <- nc_health_data_final %>%
  left_join(metro_row, by = "County") %>%
  # drop Metropolitan_or_Nonmetropolitan column
  select(-Metropolitan_or_Nonmetropolitan)

# Check it
str(nc_health_data_merged)
```

```
## tibble [99 x 17] (S3: tbl_df/tbl/data.frame)
##  $ County                                        : chr [1:99] "Alamance" "Alexander" "Alleghany" "An
##  $ ColorectalRate_I                              : num [1:99] 40.1 38.8 48 47.8 34.2 27.3 45.4 59.3
##  $ LungRate_I                                    : num [1:99] 64.7 65.8 74.1 68.4 73 65 70.5 69.5 84
##  $ BreastRate_I                                  : num [1:99] 183 177 192 204 209 ...
##  $ ColorectalRate_M                              : num [1:99] 13.1 14.3 NA 20.2 8.8 NA 15.5 17.6 17
##  $ LungRate_M                                    : num [1:99] 43.5 43.7 37.4 43.9 36.5 27.6 39.7 41
##  $ BreastRate_M                                  : num [1:99] 20.4 19.6 36.7 29 22.5 21.2 19.6 NA 28
##  $ Population                                    : num [1:99] 0.38 -0.396 -0.551 -0.473 -0.457 ...
##  $ College_Graduation                            : num [1:99] 0.172 -0.772 -0.268 -1.265 -0.184 ...
##  $ Uninsured_Adults                              : num [1:99] -0.0433 -0.6154 1.3512 -0.1148 0.4573
##  $ Health_Care_Workforce_-_Primary_Care_Physicians: num [1:99] 0.239 -0.968 -0.283 -1.327 -0.185 ...
##  $ Adult_Smoking                                 : num [1:99] -0.249 -0.249 -0.249 1.245 -0.249 ...
```

```
##  $ Poverty                                          : num [1:99] -0.253 -0.863 0.231 1.178 -0.253 ...
##  $ Food_Insecurity                                  : num [1:99] -0.471 -0.471 1.173 0.597 -0.307 ...
##  $ Transportation                                   : num [1:99] -0.447 -1.13 -0.959 1.601 -0.234 ...
##  $ Air_Pollution                                    : num [1:99] 0.92 0.472 -1.097 0.584 -1.321 ...
##  $ Metro_Status                                     : chr [1:99] "Metro" "Metro" "Nonmetro" "Nonmetro"
```

```
head(nc_health_data_merged)
```

```
## # A tibble: 6 x 17
##   County    ColorectalRate_I LungRate_I BreastRate_I ColorectalRate_M LungRate_M
##   <chr>                <dbl>      <dbl>        <dbl>            <dbl>      <dbl>
## 1 Alamance              40.1       64.7         183.             13.1       43.5
## 2 Alexander             38.8       65.8         177.             14.3       43.7
## 3 Alleghany             48         74.1         192.             NA         37.4
## 4 Anson                 47.8       68.4         204.             20.2       43.9
## 5 Ashe                  34.2       73           209.              8.8       36.5
## 6 Avery                 27.3       65           171.             NA         27.6
## # i 11 more variables: BreastRate_M <dbl>, Population <dbl>,
## #   College_Graduation <dbl>, Uninsured_Adults <dbl>,
## #   `Health_Care_Workforce_-_Primary_Care_Physicians` <dbl>,
## #   Adult_Smoking <dbl>, Poverty <dbl>, Food_Insecurity <dbl>,
## #   Transportation <dbl>, Air_Pollution <dbl>, Metro_Status <chr>
```

```
# load libraries
library(dplyr)
library(readr)
library(tidyr)
library(tidyverse)
library(broom)

# prepare the long data (keep individual county rows)
cancer_stats_raw <- nc_health_data_merged %>%
  pivot_longer(
    cols = c(ColorectalRate_I, LungRate_I, BreastRate_I,
             ColorectalRate_M, LungRate_M, BreastRate_M),
    names_to = "Metric",
    values_to = "Rate"
  ) %>%
  separate(Metric, into = c("Cancer", "Outcome"), sep = "Rate_") %>%
  mutate(
    Outcome = ifelse(Outcome == "I", "Incidence", "Mortality"),
    Metro_Status = as.factor(Metro_Status)
  ) %>%
  filter(!is.na(Metro_Status), !is.na(Rate))

# run the tests and build the summary table
results_table <- cancer_stats_raw %>%
  group_by(Cancer, Outcome) %>%
  reframe(
    # perform the t-test
    t_results = list(t.test(Rate ~ Metro_Status, var.equal = FALSE)),

    # calculate means and n for Metro
```

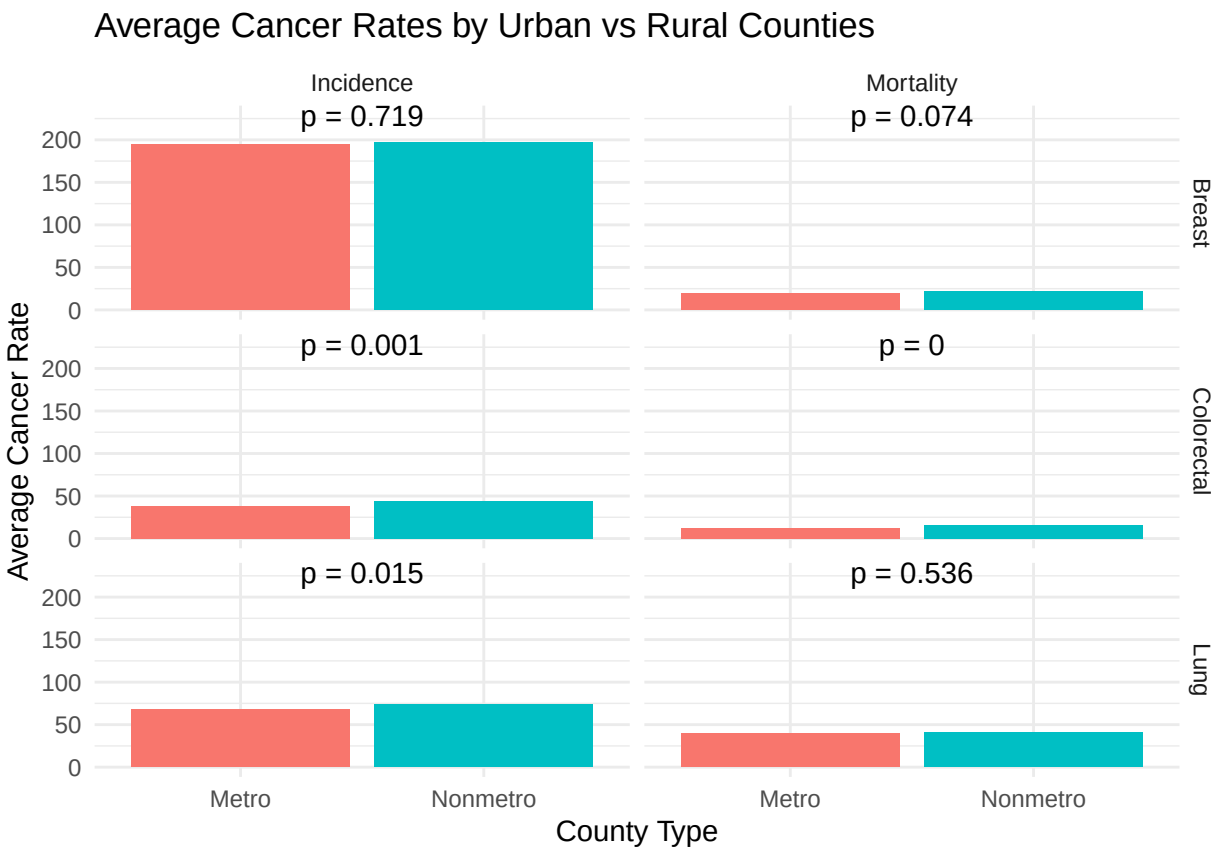
```
    Mean_Metro = mean(Rate[Metro_Status == "Metro"], na.rm = TRUE),
    n_Metro = sum(Metro_Status == "Metro"),

    # calculate means and n for Nonmetro
    Mean_Nonmetro = mean(Rate[Metro_Status == "Nonmetro"], na.rm = TRUE),
    n_Nonmetro = sum(Metro_Status == "Nonmetro")
  ) %>%
  # extract the specific stats from the t-test object
  mutate(
    t_stat = sapply(t_results, function(x) x$statistic),
    p_value = sapply(t_results, function(x) x$p.value),
    Diff = Mean_Nonmetro - Mean_Metro,
    Significant = ifelse(p_value < 0.05, "Yes", "No")
  ) %>%
  select(-t_results) %>%
  mutate(across(where(is.numeric), ~ round(., 3)))

# show results
print(results_table)
```

```
## # A tibble: 6 x 10
##   Cancer     Outcome  Mean_Metro n_Metro Mean_Nonmetro n_Nonmetro t_stat p_value
##   <chr>      <chr>         <dbl>   <dbl>         <dbl>      <dbl>  <dbl>   <dbl>
## 1 Breast     Inciden~      195.       45         197.          53 -0.361   0.719
## 2 Breast     Mortali~       19.4      27          22.0         19 -1.86    0.074
## 3 Colorectal Inciden~       38.7      45          44.1         52 -3.59    0.001
## 4 Colorectal Mortali~       12.7      42          15.7         48 -4.92    0
## 5 Lung       Inciden~       67.6      45          73.9         52 -2.47    0.015
## 6 Lung       Mortali~       39.8      45          40.8         53 -0.621   0.536
## # i 2 more variables: Diff <dbl>, Significant <chr>
```

```
# bar charts with p-value
ggplot(cancer_stats_raw, aes(x = Metro_Status, y = Rate, fill = Metro_Status)) +
  geom_bar(stat = "summary", fun = "mean", position = "dodge") +
  facet_grid(Cancer ~ Outcome) +
  labs(title = "Average Cancer Rates by Urban vs Rural Counties",
       x = "County Type",
       y = "Average Cancer Rate") +
  theme_minimal() +
  theme(legend.position = "none") +
  geom_text(data = results_table,
            aes(x = 1.5, y = max(cancer_stats_raw$Rate) * 0.9,
                label = paste0("p = ", p_value)),
            inherit.aes = FALSE)
```


Average Cancer Rates by Urban vs Rural Counties
Incidence
Mortality
p = 0.719
p = 0.074
p = 0.001
p = 0
p = 0.015
p = 0.536
Breast
Colorectal
Lung
Average Cancer Rate
0
50
100
150
200
Metro
Nonmetro
County Type