Problem Set 3: Fertility in South Korea and Across the World
Due 11:59 PM, Thursday, May 8

In this problem set, you will analyze data from the UN on fertility by country, year, and age. You will focus especially on South Korea, which experienced very rapid fertility decline in the mid-20th century. The template `Rmd` script loads a dataset that contains age-specific fertility rates by country, year, and age since 1950. Each row is a country-by-age-by-year combination.

Many questions ask you to create tables or graphs. When asked to make quantitative statements about a graph, it is fine to approximate based on looking at the graph.

1. Plot age specific fertility rates by age in South Korea (called "Republic of Korea" in the dataset) in $1950, 1960, \cdots, 2020$. How would you describe fertility change between 1960 and 1990? How about between 1990 and 2020?

2. Compute the total fertility rate for all country-year combinations in the data. Plot histograms of the total fertility rate in the years $1950, 1960, \cdots, 2020$. Describe changes in the distribution of fertility across countries.

3. Create two tables. The first table should report the global median total fertility rate in the years $1950, 1960, \cdots, 2020$. The second table should report South Korea's total fertility rate in those same years. How did South Korea's position in the global distribution of fertility change from 1950 to 2020?

4. What is the relationship between late teen fertility and the total fertility rate? Draw a separate scatterplot for the each year $1950, 1960, \cdots, 2020$. Your scatterplots should have TFR on the y-axis and $ASFR_{19}$ on the x-axis.[1] Describe the relationship, and assess whether it changes over time.

5. Repeat the previous exercise for South Korea by itself. Now you should draw a single scatterplot where each point is a separate year. It should still have TFR on the y-axis and $ASFR_{19}$ on the x-axis. Use all years, not just those ending in 0. Does the Korea-only scatterplot over time look similar to the cross-country scatterplots from Question 4? If it is different, how so?

6. So far, you have taken a period approach to studying fertility in South Korea. Now switch to a cohort approach. As a first step to the cohort approach, create a data frame for South Korea in which each row corresponds to a birth year and age. Keep only birth years ending in 0. Create a table that reports how many observations you have for each birth year.[2] For which birth years do you have a full fertility history?

[1] You may find that you want to join data frames by multiple variables. See the methods lecture from week 2 for an example of joining data frames. To join data frames `df1` and `df2` by the variables `x` and `y`, you can type: `df1 |> left_join(df2, by = c("x", "y"))`.

[2] Hint: in `tidyverse`, you can use `group_by()` and `summarise(count = n())`.

7. For the cohorts you identified as having full fertility histories in Question 6, plot $ASFR$ against age by cohort. Between which cohorts was there a decrease in $ASFR$ at almost all ages? Between which cohorts was there a shift toward later childbearing?

8. Compute children ever born at each age for the cohorts in your graph for Question 7. Plot children ever born against age by cohort. How much did average children ever born at age 30 decline between the 1940 cohort and the 1970 cohort? How about the completed fertility rate?

9. Compute the completed fertility rate for every Korean cohort with a full fertility history. Plot the completed fertility rate against birth year for all relevant cohorts. When did most of South Korea's fertility decline occur?

10. The risk of chromosomal abnormalities like Down Syndrome is higher for babies born to women in their 40s. Has the rate of giving birth at age 40 been rising or falling across cohorts in South Korea? Plot the evolution of this rate across cohorts, and describe your results.