ECON 125 Problem Set 3 Solution

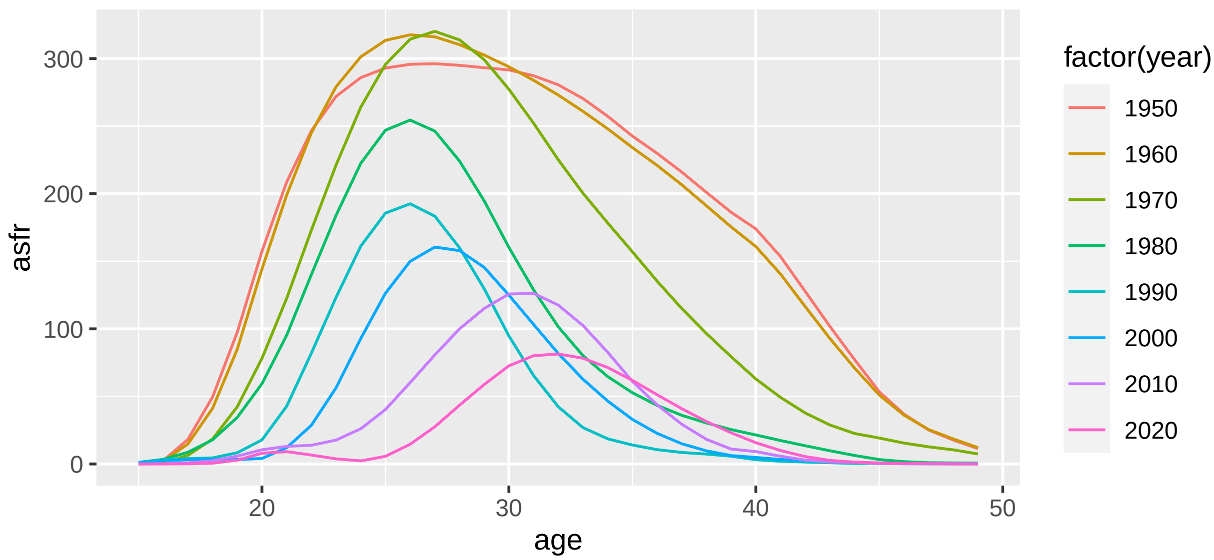
Question 1

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?

Answer: To plot South Korean ASFRs in years ending in 0, we split off a data frame and then plot using that data frame.

```
# Split off korea data frame
korea <-
  country_year_age |>
  filter(country=="Republic of Korea"&year%%10==0)

# Plot rates by age and year
ggplot(korea, aes(x=age, y=asfr, color=factor(year))) +
  geom_line()
```



We see fertility rates fell dramatically between 1960 and 2020 but also changed differentially across ages. Between 1960 and 1990, peak fertility remained stable in the mid-20s, but as fertility fell, it became increasingly concentrated at those ages. Between 1990 and 2020, peak fertility shifted later in the life cycle, from the mid-20s to the early-30s. Peak age-specific fertility in 1960 was about four times that in 2020.

Question 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.

Answer: We first compute the TFR as the sum of age-specific rates. We then split off the country-decade data frame by filtering to years ending in 0. Finally, we plot histograms by year using the `facet_wrap()` syntax we learned in the Week 1 methods slides.

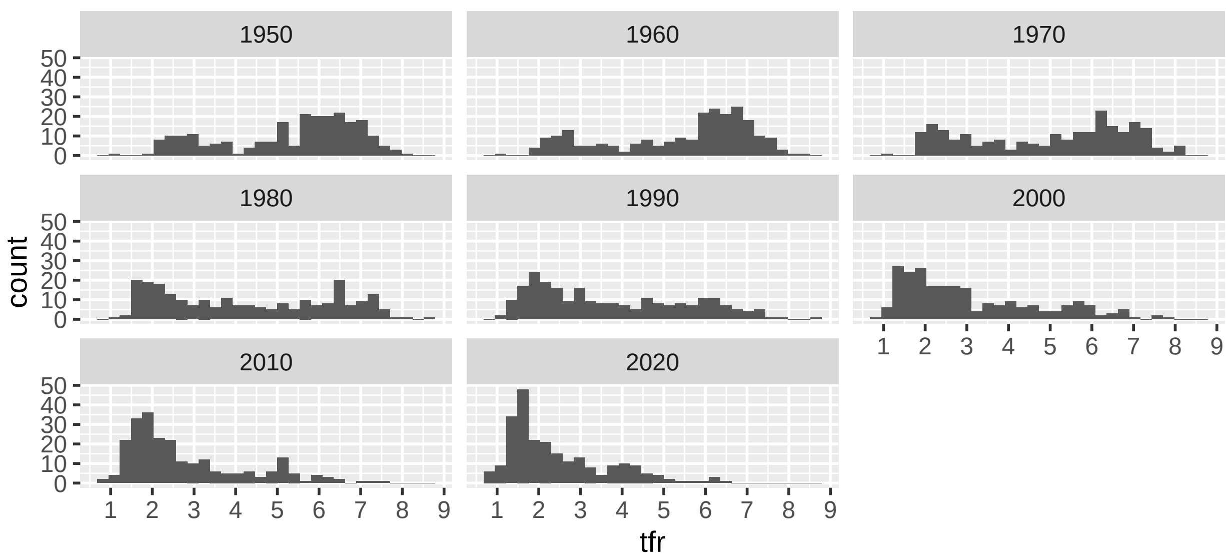
```
# Create a country-year data frame with total fertility rates
country_year <-
  country_year_age |>
  group_by(country, year) |>
  summarise(tfr = sum(asfr)/1000)
```

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

```
# Split off a data frame only containing years ending in 0
country_decade <- country_year |> filter(year%%10==0)

# Plot histograms for every tenth year using facet_wrap.
# I modified the x-axis grid to have integer intervals,
# but this was not necessary.
ggplot(country_decade, aes(x = tfr)) +
  geom_histogram() +
  scale_x_continuous(breaks = 0:9) +
  facet_wrap(~year)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



In the 1950s and 1960s, many countries had TFRs above 5, and vanishingly few countries had TFRs below 2. By 2020, the distribution had shifted substantially. In that year, vanishingly few countries had TFRs above 5, and the mode of the distribution was below 2.

Question 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?

Answer: We can use the `country_decade` data frame from the last problem. First, we group by year and summarize the median. Second, we filter to South Korea.

```
# Table 1: Median every tenth year
country_decade |>
  group_by(year) |>
  summarise(med = median(tfr))
```

```
## # A tibble: 8 x 2
##    year   med
##   <dbl> <dbl>
## 1  1950  5.77
## 2  1960  6.00
## 3  1970  5.41
## 4  1980  4.02
## 5  1990  3.26
## 6  2000  2.57
## 7  2010  2.25
## 8  2020  2.02
```

```
# Table 2: Restrict country_decade to South Korea
country_decade |>
  filter(country=="Republic of Korea")
```

```
## # A tibble: 8 x 3
## # Groups:   country [1]
##   country            year   tfr
##   <chr>             <dbl> <dbl>
## 1 Republic of Korea  1950 6.05
## 2 Republic of Korea  1960 5.99
## 3 Republic of Korea  1970 4.45
## 4 Republic of Korea  1980 2.73
## 5 Republic of Korea  1990 1.60
## 6 Republic of Korea  2000 1.47
## 7 Republic of Korea  2010 1.23
## 8 Republic of Korea  2020 0.812
```

South Korea's TFR fell from 6.1 in 1950 to 0.8 in 2020. The TFR of 6.1 made it a fairly typical country in 1950, slightly above the cross-country median of 5.8. The $TFR of 0.8 made it a very low fertility country in 2020, well below the median of 2.0.

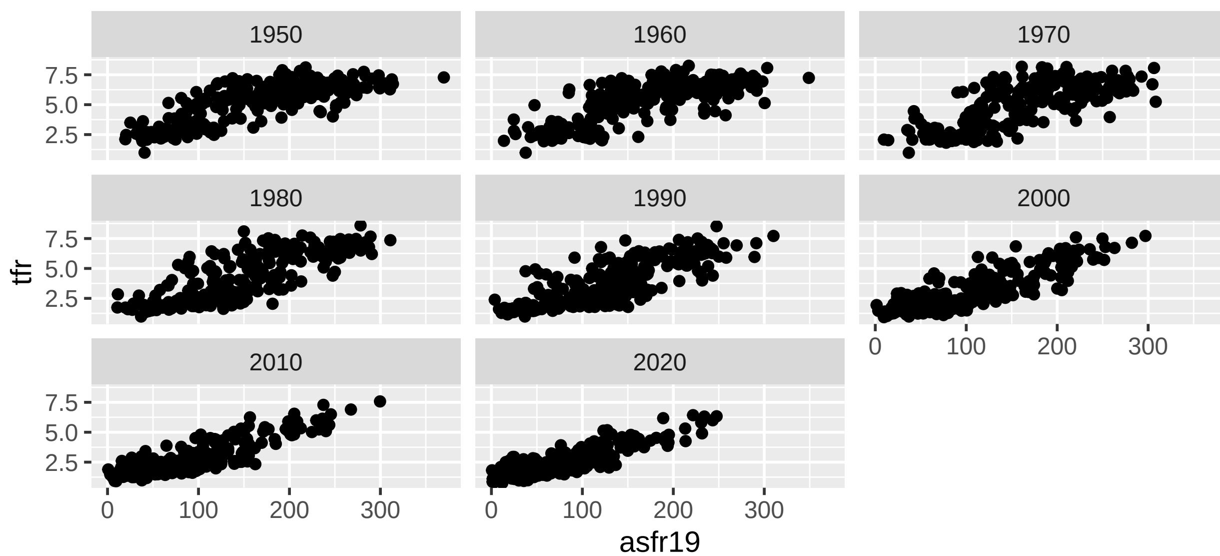
Question 4

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.

Answer: We will restrict `country_year_age` to only contain age 19 fertility, and then join it with `country_decade`, which already has TFR for every tenth year. Then we will draw separate scatterplots for each year using `facet_wrap()`.

```r
# Restrict country_year_age to ASFR19,
# join it with country_decade,
# and write over country_decade.
country_decade <-
  country_year_age |>
  filter(year%%10==0 & age==19) |>
  select(country, year, asfr) |>
  rename(asfr19 = asfr) |>
  left_join(country_decade, by = c("country", "year"))

# Draw the plot
ggplot(country_decade, aes(x=asfr19, y=tfr)) +
  geom_point() +
  facet_wrap(~year)
```



There is a positive, linear relationship between age-19 fertility and the total fertility rate. The slope is fairly stable over time, such that an increase in age-19 fertility of 100 births per 1000 women is associated with a roughly 2 child per woman increase in the total fertility rate.

[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"))`.

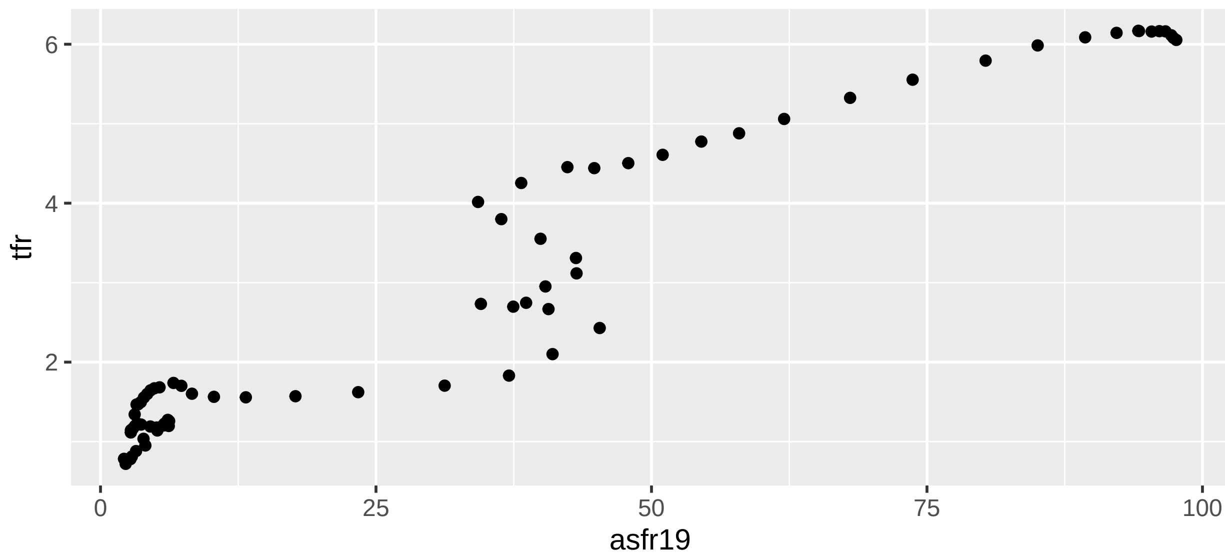
Question 5

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?

Answer: We will use the same coding approach as in Question 4.

```
# Create a Korea-only data frame with tfr and asfr19 for every year
korea <-
  country_year_age |>
  filter(age==19) |>
  select(country, year, asfr) |>
  rename(asfr19 = asfr) |>
  left_join(country_year, by = c("country", "year")) |>
  filter(country=="Republic of Korea")

# Draw the plot
ggplot(korea, aes(x=asfr19, y=tfr)) +
  geom_point()
```



The within-Korea relationship between teenage fertility and the total fertility rate is *very* different from the cross-country relationship. The latter was linear, with a slope of approximately 0.02. Here, the overall slope from comparing the highest point to the lowest point is approximately 0.06: three times larger! But more importantly, the relationship is not linear. The TFR fell from 4 to 2 with hardly any change in $ASFR_{19}$. And then $ASFR_{19}$ fell from 75 to 10 with hardly any change in the TFR. Teenage fertility is less systematically linked the TFR within Korea than across countries.

Question 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?

Answer: We write over the existing `korea` data frame and then create the table.

```
# New cohort data frame for Korea
korea <-
  country_year_age |>
  mutate(birthyear = year-age) |>
  select(country, birthyear, age, asfr) |>
  filter(country=="Republic of Korea" & birthyear%%10==0)

# Histogram of birth year
korea |>
  group_by(birthyear) |>
  summarise(count = n())
```

```
## # A tibble: 10 x 2
##    birthyear count
##        <dbl> <int>
##  1      1910    10
##  2      1920    20
##  3      1930    30
##  4      1940    35
##  5      1950    35
##  6      1960    35
##  7      1970    35
##  8      1980    29
##  9      1990    19
## 10      2000     9
```

We have a full fertility history for 1940-1970.

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

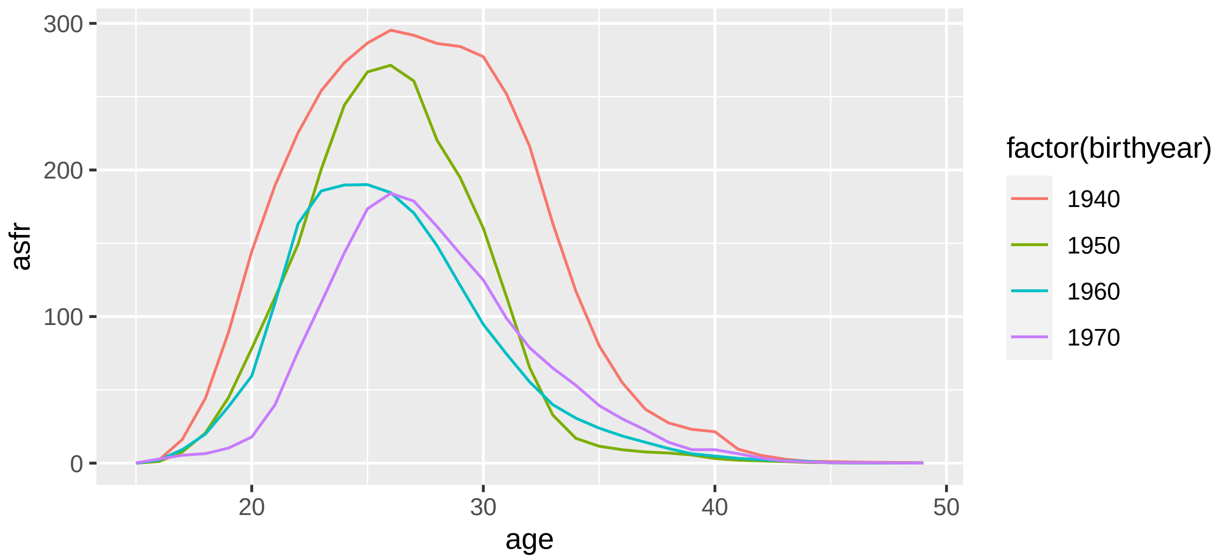
Question 7

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?

Answer: We restrict the `korea` data frame to the relevant cohorts and then draw the plot.

```
# Restrict to cohorts with complete histories
korea <- korea |> filter(birthyear>=1940 & birthyear<=1970)

# Draw the plot
ggplot(korea, aes(x = age, y = asfr, color = factor(birthyear))) +
  geom_line()
```



There was an across-the-board decrease in $ASFR$s between the 1940 and 1950 cohorts. There was a shift toward later childbearing between the 1960 and 1970 cohorts.

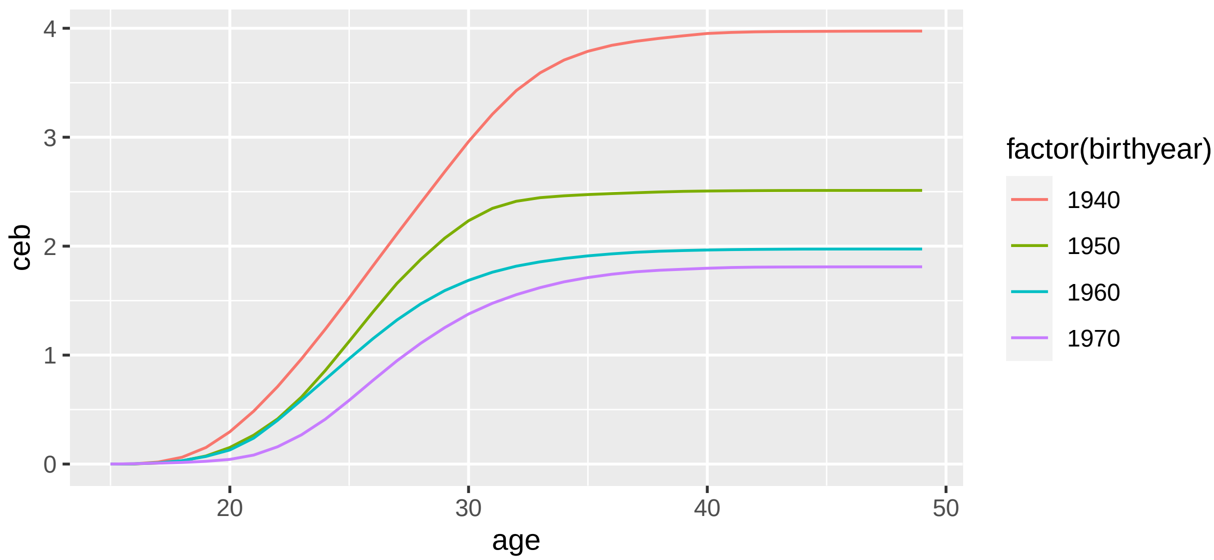
Question 8

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?

Answer: As we saw in the methods lecture from week 4, we can obtain average children ever born by taking the cumulative sum of $ASFRs$ within the cohort and dividing by 1000.

```
# Generate children ever born at each age
korea <-
  korea |>
  group_by(birthyear) |>
  arrange(age) |>
  mutate(ceb = cumsum(asfr)/1000)

# Draw plot
ggplot(korea, aes(x=age, y=ceb, color=factor(birthyear))) +
  geom_line()
```



Children ever born at age 30 fell by slightly more than 1.5 between the 1940 cohort and the 1970 cohort. The completed fertility rate fell by slightly more than 2.

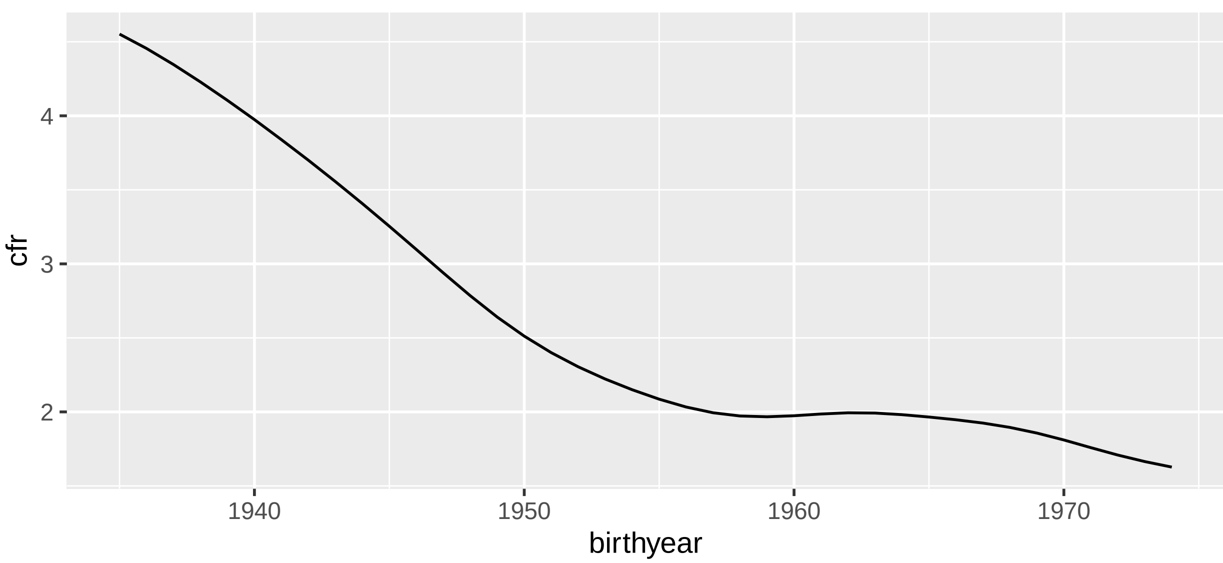
Question 9

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?

Answer: We reuse and modify code from above to form the data frame with more birth years.

```
# New version of korea data frame with CFR for all birth years with complete histories
korea <-
  country_year_age |>
  mutate(birthyear = year-age) |>
  filter(country=="Republic of Korea") |>
  select(birthyear, age, asfr) |>
  group_by(birthyear) |>
  arrange(age) |>
  summarise(cfr = sum(asfr)/1000,
            count = n()) |>
  filter(count==35)

# Draw plot
ggplot(korea, aes(x=birthyear, y=cfr)) +
  geom_line()
```



Most fertility decline happened across the first half of the cohorts included in the sample. The CFR declined from roughly $4\frac{1}{2}$ to roughly 2 between the cohorts of 1935 and 1955. It was then flat for a decade and then gradually declined to 1.6 between the cohorts of the mid-1960s and mid-1970s.

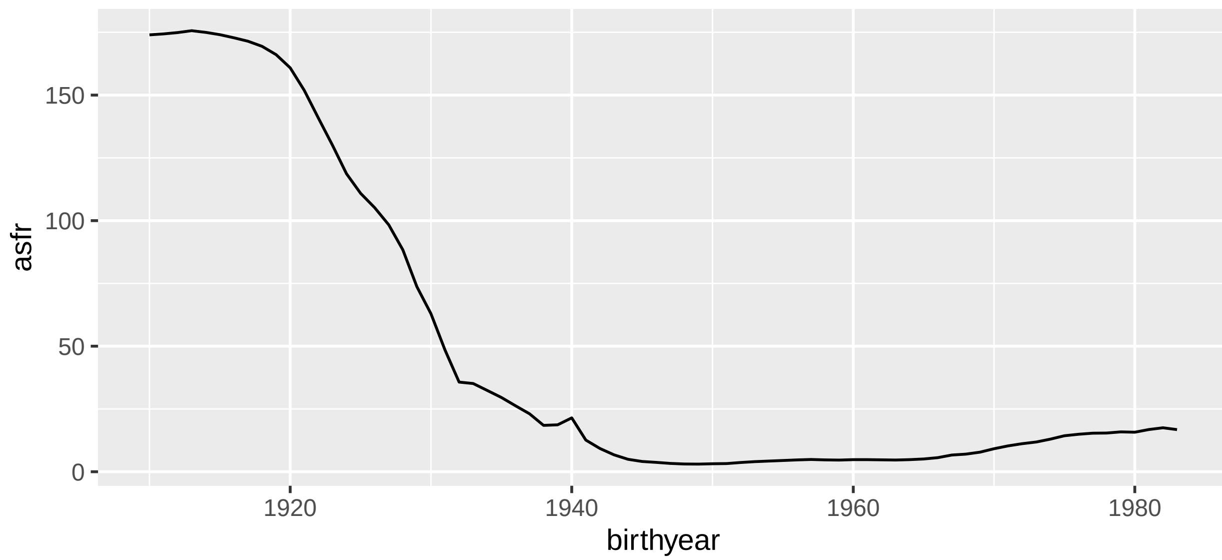
Question 10

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.

Answer: We write over the `korea` data frame again, this time with age 40 $ASFR$s.

```
# New version of korea data frame with ASFR40
korea <-
  country_year_age |>
  mutate(birthyear = year-age) |>
  filter(country=="Republic of Korea" & age==40) |>
  select(birthyear, asfr)

# Draw plot
ggplot(korea, aes(x=birthyear, y=asfr)) +
  geom_line()
```



The rate of giving birth at age 40 declined by a factor of 50 between cohorts of the 1910s and cohorts of the late 1940s, reaching a low of 3. However, it has been rising steadily since the cohorts of the 1960s, and it now exceeds 15. We can therefore predict that risk of chormosomal abnormalities has been rising in recent cohorts.