Week 4 Methods: Fertility

ECON 125: The Science of Population

Setup

Today, we analyze fertility over time within the United States

We will analyze two separate datasets: first for the full population, and second for specific racial and ethnic groups

Our first dataset includes:

- Births and mid-year female population
- One row for each single year of age (12-55) for each year (1933-2022)
- From the Human Fertility Database, based on US Vital Statistics

Start by setting up R and loading the first dataset

```
# Load tidyverse and clear the R environment
library(tidyverse)
rm(list=ls())

# Load dataset
age_year_df <- read_csv(url("https://github.com/tomvogl/econ125/raw/mai
```

Variables

Let's look at the first few rows of the dataset

```r
head(age_year_df, 3)
```

```
## # A tibble: 3 x 4
##    year   age births   women
##   <dbl> <dbl>  <dbl>   <dbl>
## 1  1933    12     47 1237992
## 2  1933    13    532 1216487
## 3  1933    14   2121 1197058
```

As we did in the mortality data exercise, let's rename age as `x` in the data

```r
age_year_df <- age_year_df |> rename(x = age)
```

Our building block for today is the **age-specific fertility rate** at age x

$$ASFR_x = 1000 \times \frac{births_x}{women_x}$$

```r
age_year_df <- age_year_df |> mutate(asfr = 1000*births/women)
```

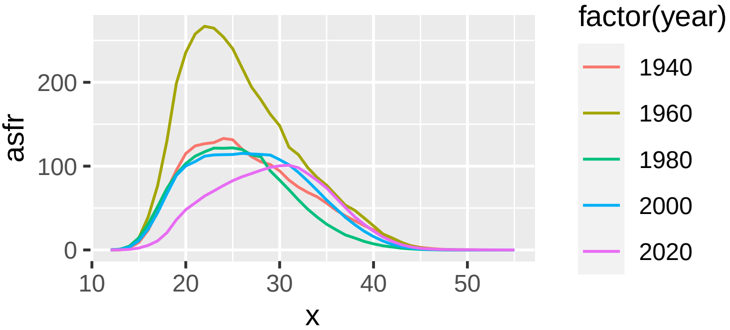
Age-specific fertility rates over time

Let's split off a data frame with every 20^{th} year

```
twenty_df <- age_year_df |> filter(year%%20==0)
```

Now let's plot the age pattern of fertility by year

```
ggplot(twenty_df, aes(x = x, y = asfr, color = factor(year))) +
  geom_line()
```



Age-specific fertility rates are highest in 20s and 30s

Can very easily see the Baby Boom in 1960 and delayed fertility in 2020

Crude birth rate: definition

As with mortality, we often want a single measure to describe the level of fertility

The simplest measure is the crude birth rate

- *CBR* equals total births divided by total population (usually $\times 1000$)

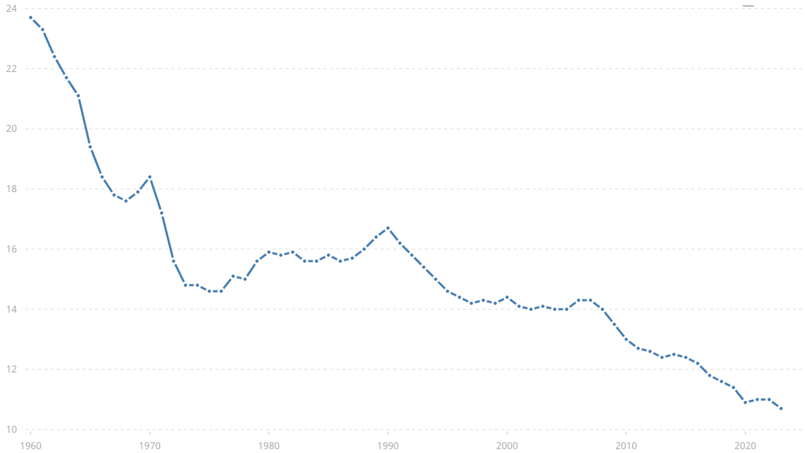
$$CBR = 1000 \times \frac{births}{population}$$

- Low information requirement: only births and people, no age
- Dataset has # women of childbearing age but not # people
- So rather than compute *CBR* ourselves, let's look at a graph from elsewhere

Crude birth rate: time series

From the World Bank's *World Development Indicators* website

Crude birth rates for the United States (UN data)



General fertility rate: definition

A slightly more refined measure is the general fertility rate

- *GFR* equals total births divided by total women of childbearing age (15-44)

$$GFR = 1000 \times \frac{births}{women \text{ 15-44}}$$

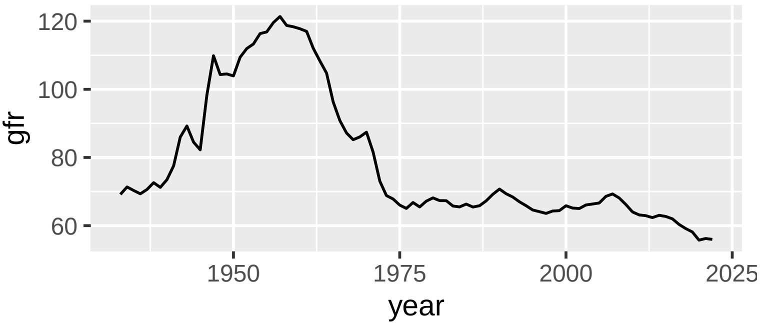
- Slightly higher information requirement than *CBR*: need counts of women by age, but not mothers' ages at birth
- Can calculate in our dataset

```
year_df <-
  age_year_df |>
  mutate(women_childbearing = if_else(x>=15 & x<45, women, 0)) |>
  group_by(year) |>
  summarise(gfr = 1000*sum(births)/sum(women_childbearing))
```

General fertility rate: time series

Let's look at the time series of the *GFR*

```
ggplot(year_df, aes(x = year, y = gfr)) +
  geom_line()
```



Compared with the *CBR* series, the *GFR* series is:

- Similar after the Baby Boom, with large decline 1960-75
- Flatter 1990-2010 because GFR omits the growing elderly from denominator

CBR versus *GFR*

CBR is very susceptible to variation in age structure

- Populations with large elderly shares will tend to have low *CBR*
- *CBR* still useful $\rightarrow$ directly contributes to the population growth rate
- Recall that in a closed population (no migration), $growth = CBR - CMR$

CBR versus *GFR*

CBR is very susceptible to variation in age structure

- Populations with large elderly shares will tend to have low *CBR*
- *CBR* still useful $\rightarrow$ directly contributes to the population growth rate
- Recall that in a closed population (no migration), $growth = CBR - CMR$

GFR fixes the sensitivity to age structure to some extent

- Removes men, children, and elderly from the denominator
- Still, age structure within 15-44 could vary across populations
- In principle, could calculate age-standardized fertility rate, but uncommon

Total fertility rate

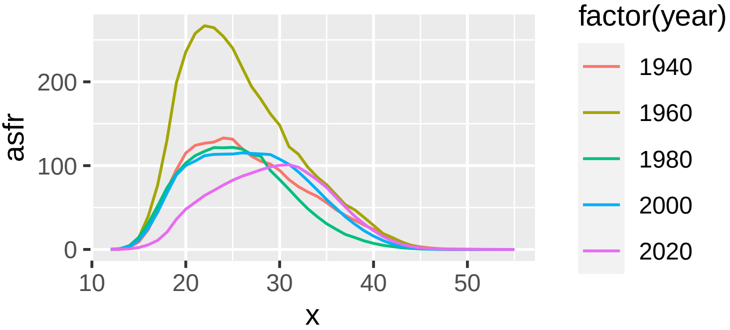
The more common way to address concerns about age structure is the *TFR*

TFR is a similar thought experiment to period life expectancy

Asks: *If a woman experienced today's age-specific fertility rates at every age, how many children could she expect to have over her life?*

Despite similarity to period life expectancy, *TFR* is **much** easier to calculate

Recall the *ASFR* graph → *TFR* is simply the sum of these rates across all ages



Total fertility rate: definition

With single-year age intervals, the total fertility rate is simply:

$$TFR = \sum_x ASFR_x / 1000$$

Typically, x runs from 15 to 44 or from 15-49

In our data, we have ages 12-55, but with few births before 12 and after 45, so it makes little difference

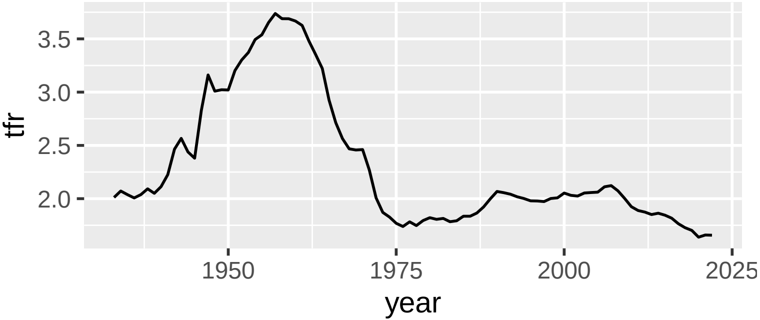
Calculation in R is simple

```
year_df <-
  age_year_df |>
  group_by(year) |>
  summarise(tfr = sum(asfr/1000))
```

Total fertility rate: time series

Let's look at the time series of the *TFR*

```
ggplot(year_df, aes(x = year, y = tfr)) +
  geom_line()
```



Much easier to interpret *TFR* than *CBR* and *GFR*! Children per woman...

- 2 in 1930s
- Almost 4 circa 1960
- At or below 2 since early 1970s

Thinking about timing

TFR was 3.7 in late 1960 and 1.6 in 2020

Does this mean that women in the 1960s had 3.7 children on average?

No!

Thinking about timing

TFR was 3.7 in late 1960 and 1.6 in 2020

Does this mean that women in the 1960s had 3.7 children on average?

No!

Timing matters

- Possible for *TFR* to swing wildly, even with a constant lifetime # children
- For example, suppose all women have exactly 2 children over their lifetimes
- Suppose gov't announces it will pay $1m to every pregnant woman this year
- *TFR* will surge this year and drop next year, but lifetime fertility constant

Quantum versus tempo

Quantum effect
Refers to **total number of children** a cohort of women have over their lifetimes
- Example: A drop in completed fertility from 2.1 to 1.8 children per woman

Tempo effect
Refers to **shifts in the timing** of childbearing
- Example: Women delay first birth $\rightarrow$ *TFR* falls but lifetime fertility unchanged

Cohort fertility

To learn about the quantum of fertility, need to switch to a cohort perspective

Can approximate cohort fertility by stringing together period age-specific rates

- E.g., $ASFR_{15}^{1990}, ASFR_{16}^{1991}, \cdots, ASFR_{44}^{2019}$
- "Approximate" because $ASFR_{15}^{1990}$ mixes women born in 1974 and 1975
- But still very close to what we would get from following the 1975 cohort

Create age-cohort data frame for ages 15-45

```
age_cohort_df <-
  age_year_df |>
  filter(x>=15 & x<=45) |>
  mutate(birthyear = year - x)
```

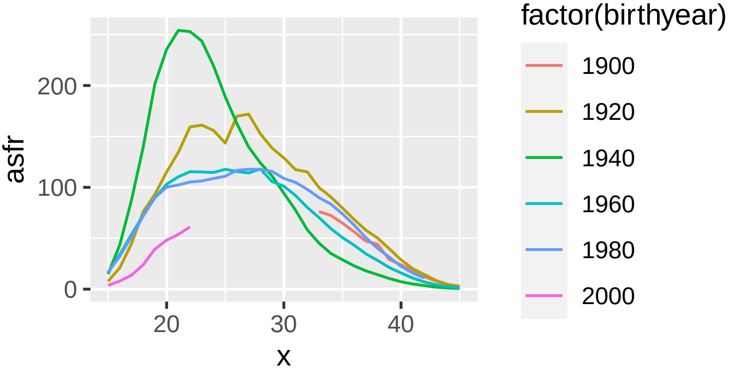
Plot cohort $ASFR_x$ by age

Let's split off a data frame with every 20^{th} cohort

```
twenty_df <- age_cohort_df |> filter(birthyear%%20==0)
```

And then draw the evolution of age-specific fertility for specific cohorts

```
ggplot(twenty_df, aes(x = x, y = asfr, color=factor(birthyear))) +
  geom_line()
```



Translating age-specific rates to children ever born

Let's compute children ever born at age x (CEB_x) for each cohort and age

To do so, we need to focus on cohorts that we can see starting at age 15

```
age_cohort_df <-
  age_cohort_df |>
  group_by(birthyear) |>
  mutate(min_age = min(x)) |>
  filter(min_age==15)
```

And then compute the running sum of age-specific rates

```
age_cohort_df <-
  age_cohort_df |>
  group_by(birthyear) |>
  arrange(x) |>
  mutate(ceb = cumsum(asfr/1000))
```

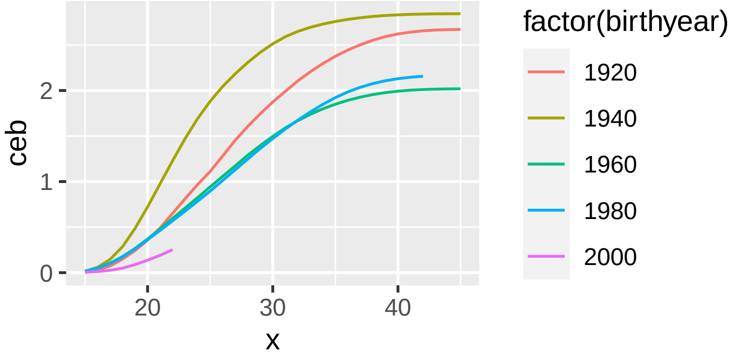
Lifecycle children ever born by cohort

Let's split off a data frame with every 20^{th} cohort again

```
twenty_df <- age_cohort_df |> filter(birthyear%%20==0)
```

And then draw the evolution of CEB_x

```
ggplot(twenty_df, aes(x = x, y = ceb, color=factor(birthyear))) +
  geom_line()
```



1920 and 1940 cohorts had similar lifetime numbers, but 1940 had them earlier

Completed fertility rate

The completed fertility rate is children ever born at end of reproductive period

$$CFR = CEB_{45}$$

Easy to obtain from our existing data frame: just keep observations with $x = 45$

```
cohort_df <-
  age_cohort_df |>
  filter(x==45) |>
  rename(cfr = ceb) |>
  select(birthyear, cfr)
```

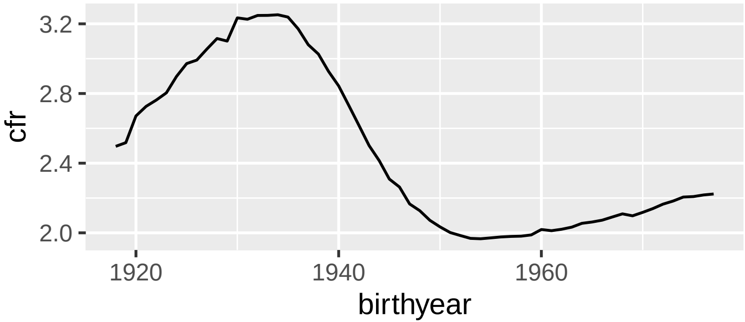
What is the distribution of *CFR* across cohorts?

```
summary(cohort_df)
```

```
##    birthyear         cfr
##  Min.   :1918   Min.   :1.966
##  1st Qu.:1933   1st Qu.:2.060
##  Median :1948   Median :2.220
##  Mean   :1948   Mean   :2.461
##  3rd Qu.:1962   3rd Qu.:2.905
##  Max.   :1977   Max.   :3.252
```

Plotting completed fertility rate across cohorts

```r
ggplot(cohort_df, aes(x = birthyear, y = cfr)) +
  geom_line()
```



Can see that women never had more than 3.5 kids on average, nor less than 1.75

Have to wait for cohorts to age out of childbearing → cohort patterns out of date

Back to period data

Explore fertility differences across racial/ethnic groups within the US

Another dataset!

- Age-specific fertility rates for 5-year age groups (10-14 to 45-49)
- One row for each year (1993-2019) and racial/ethnic group
- From the National Center for Health Statistics, based on US Vital Statistics

```
# Clear the R environment to keep it uncluttered
rm(list=ls())

# Load dataset
race_year_df <- read_csv(url("https://github.com/tomvogl/econ125/raw/ma
```

Variables

Let's look at the first few rows of the dataset

```
head(race_year_df, 3)
```

```
## # A tibble: 3 x 10
##    year race      asfr1014 asfr1519 asfr2024 asfr2529 asfr3034 asfr3
##   <dbl> <chr>        <dbl>    <dbl>    <dbl>    <dbl>    <dbl>    <d
## 1  1993 non-hisp~      0.5     40.7     92.2     108.     79        3
## 2  1994 non-hisp~      0.5     40.4     90.9     107.     80.2      3
## 3  1995 non-hisp~      0.4     39.3     90.2     105.     81.5      3
## # i 1 more variable: asfr4549 <dbl>
```

Different structure from before: now each age group gets its own column

Researchers call this a **wide format** dataset, in contrast with **long format**

Race/ethnicity

Let's see what race/ethnicity categories are in the data

```
race_year_df |> distinct(race)
```

```
## # A tibble: 5 x 1
##   race
##   <chr>
## 1 non-hispanic white
## 2 non-hispanic black
## 3 american indian or alaska native
## 4 asian or pacific islander
## 5 hispanic
```

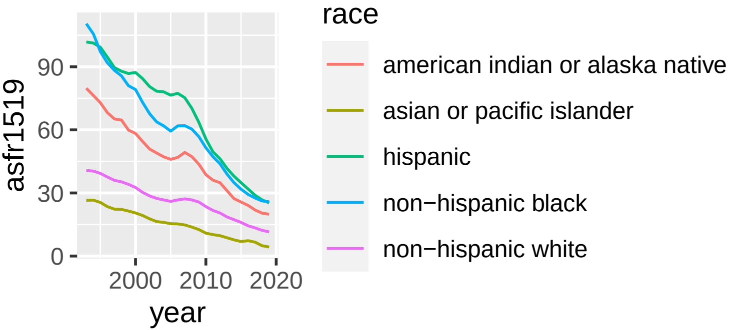
We want to calculate *TFR* for each race-year combination

Teen birth rate by race/ethnicity and year

Interesting to consider the teen birth rate by race/ethnicity over time

Many policies during this period tried to reduce the teen birth rate

```r
ggplot(race_year_df, aes(x = year, y = asfr1519, color = race)) +
  geom_line()
```



Steep decline in teen fertility, especially for disadvantaged groups

- All groups fell by $> 70\%$, largest in absolute terms for Black and Hispanic

TFR by race/ethnicity and year

Code for calculating *TFR* will be a bit different from before

- Need to sum across `asfr1014` to `asfr4549`
- Need to multiply by 5 because a woman spends five years in each age group
- No longer need `group_by()` because the observations are already grouped

```
race_year_df <-
  race_year_df |>
  mutate(tfr = 5/1000*(asfr1014 + asfr1519 + asfr2024 + asfr2529 +
                       asfr3034 + asfr3539 + asfr4044 + asfr4549))
```

Initial *TFR* by race/ethnicity

When I was a kid in the 1990s, Hispanic, Black, Native women had more kids

Hispanic women prevented the US from below-replacement fertility (<2.1)

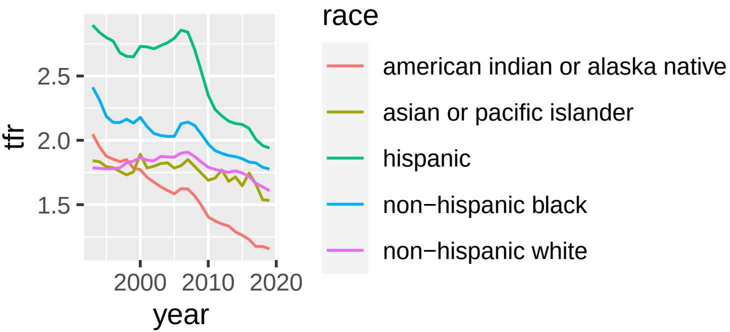
We can see this pattern by reporting *TFR* by race/ethnicity in 1993

```
race_year_df |>
  filter(year==1993) |>
  select(race, tfr) |>
  arrange(race)
```

```
## # A tibble: 5 x 2
##   race                               tfr
##   <chr>                            <dbl>
## 1 american indian or alaska native  2.05
## 2 asian or pacific islander         1.84
## 3 hispanic                          2.89
## 4 non-hispanic black                2.41
## 5 non-hispanic white                1.79
```

TFR time series by race/ethnicity

```r
ggplot(race_year_df, aes(x = year, y = tfr, color = race)) +
  geom_line()
```



For a long time, Hispanic women kept US *TFR* above replacement level

No longer! Hispanic *TFR* fell 32% in 2006-19

Native *TFR* fell 44% in 1993-2019