Week 6 Methods: Gender

ECON 125: The Science of Population

Setup

Today, we analyze sex ratios, fertility-stopping patterns, and boy-girl mortality differences in India

We use the National Family Health Survey 4, a nationally representative survey of women of childbearing age collected in 2015-26

The dataset includes:

- Mother-level and birth-level variables
- One row for every birth to mothers who responded to the survey

Start by setting up R and loading the first dataset

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

# Load dataset
nfhs4 <- read_csv(url("https://github.com/tomvogl/econ125/raw/main/data

# Ask R to not use scientific notation (not essential)
options(scipen = 999)
```

Variables

`mom_id` = mother's unique identifying number

`mom_age` = mother's age

`mom_kids` = mother's number of children ever born

`birth_order` = birth order

`birth_year` = birth year

`birth_male` = 1 if infant male, 0 if female

`birth_u1` = 1 if infant died before age 1, 0 if survived

Glimpse

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

We'll use `glimpse()` instead of `head()` because the dataset has many columns

`glimpse()` transposes them

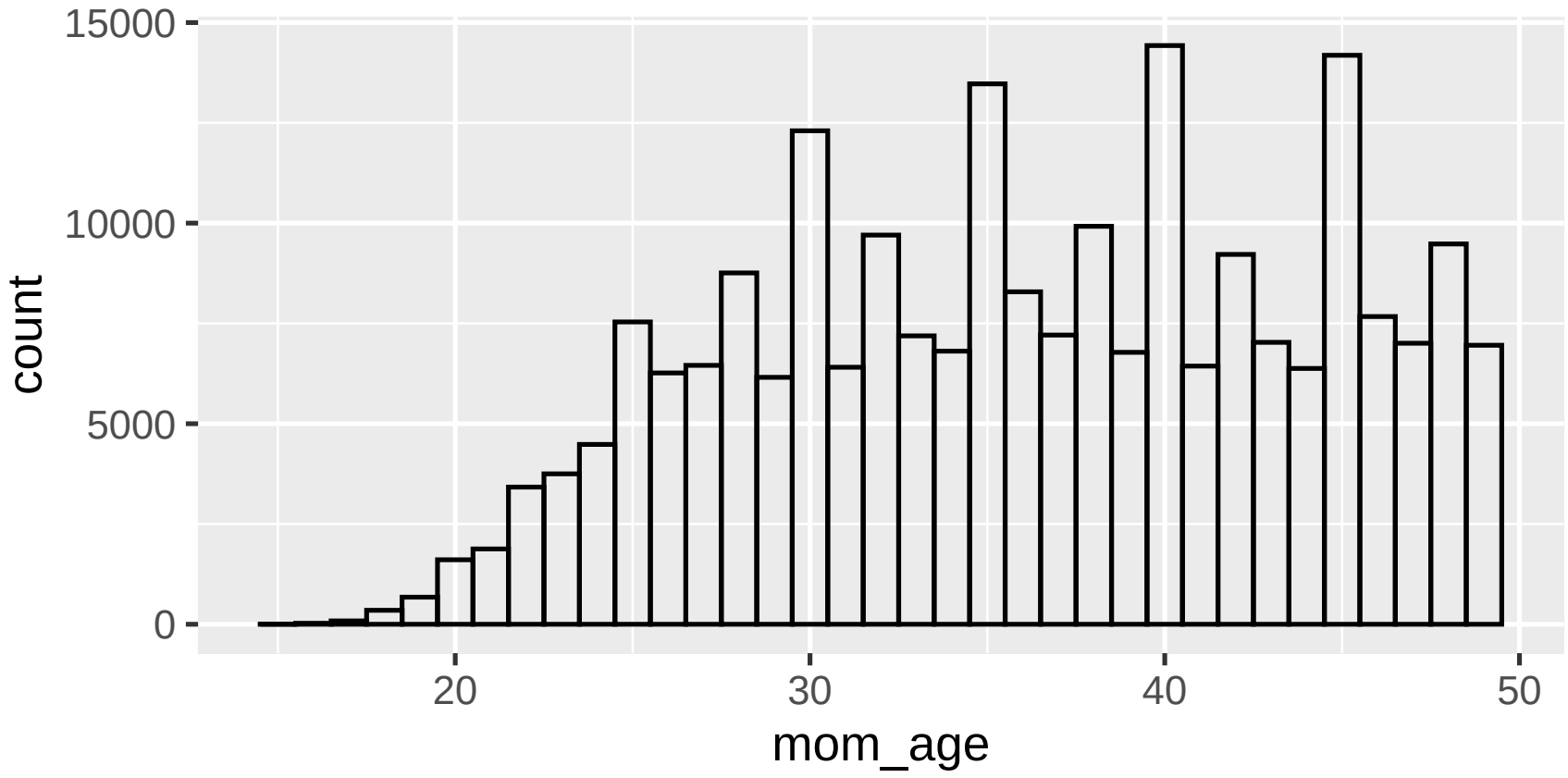
```
glimpse(nfhs4)
```

```
## Rows: 228,319
## Columns: 7
## $ mom_id      <dbl> 5, 5, 7, 7, 8, 8, 8, 10, 10, 12, 12, 13, 13, 16,
## $ mom_age     <dbl> 35, 35, 46, 46, 40, 40, 40, 40, 40, 37, 37, 33,
## $ mom_kids    <dbl> 2, 2, 2, 2, 3, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2,
## $ birth_order <dbl> 2, 1, 2, 1, 3, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2,
## $ birth_year  <dbl> 2002, 1999, 1997, 1989, 1998, 1993, 1992, 1995,
## $ birth_male  <dbl> 1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1,
## $ birth_u1    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
```

Age of moms

Consider the histogram of mom's age

```
ggplot(nfhs4, aes(x=mom_age)) +
  geom_histogram(binwidth=1, fill = NA, color = "black")
```



Few young women (why?), also note age heaping

Fertility of moms

Consider the mean number of children ever born

```
nfhs4 |> summarise(mean_kids = mean(mom_kids))
```

```
## # A tibble: 1 x 1
##   mean_kids
##       <dbl>
## 1      3.63
```

This number is much higher than India's TFR, which is around 2

- Reasons that it would overstate average fertility?
- Reasons that it would understate average fertility?

Fertility of older moms

Consider the mean number of children ever born to women over 45

Now we will keep only the last observation for each woman

```
nfhs4 |>
  filter(mom_age>=45 & birth_order==mom_kids) |>
  summarise(mean_kids = mean(mom_kids))
```

```
## # A tibble: 1 x 1
##   mean_kids
##       <dbl>
## 1      3.65
```

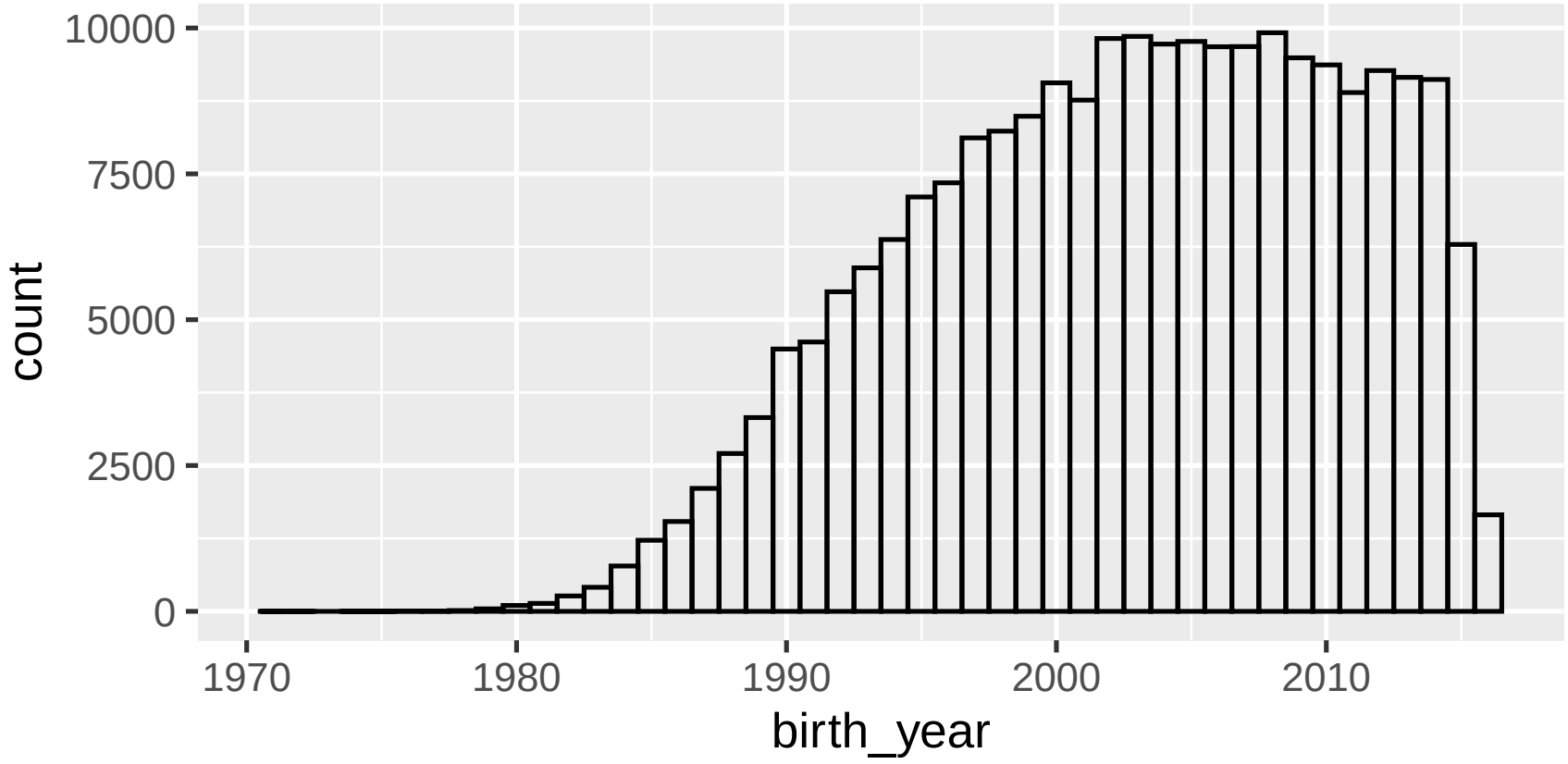
Basically the same, despite older sample!

- Fewer biases here than in the last sample
- But still not the completed fertility rate $\rightarrow$ missing childless women

Child birth years

Consider the histogram of child birth year

```
ggplot(nfhs4, aes(x=birth_year)) +
  geom_histogram(binwidth=1, fill = NA, color = "black")
```



Most births are in the 2000s and 2010s

Sex ratio

What is the overall share male at birth?

```
nfhs4 |> summarise(share_male = mean(birth_male))
```

```
## # A tibble: 1 x 1
##   share_male
##        <dbl>
## 1      0.521
```

How about the sex ratio at birth?

```
nfhs4 |> summarise(ratio = 100*mean(birth_male)/(1 - mean(birth_male)))
```

```
## # A tibble: 1 x 1
##   ratio
##   <dbl>
## 1  109.
```

The sex ratio at birth is 109 boys per 100 girls

Birth order

We will want to look at patterns by birth order

Start by tabulating birth orders 1-8

```
nfhs4 |>
  group_by(birth_order) |>
  summarise(n = n(), pct = n/nrow(nfhs4)) |>
  filter(birth_order<=8)
```

```
## # A tibble: 8 x 3
##   birth_order     n     pct
##         <dbl> <int>   <dbl>
## 1           1 83039 0.364
## 2           2 66627 0.292
## 3           3 38810 0.170
## 4           4 20354 0.0891
## 5           5 10184 0.0446
## 6           6  5056 0.0221
## 7           7  2405 0.0105
## 8           8  1063 0.00466
```

Very few kids of birth order 6+, so we will focus on 1-5

Sex ratio by birth order

How does the sex ratio depend on birth order?

```
nfhs4 |>
  filter(birth_order<6) |>
  group_by(birth_order) |>
  summarise(share_male = mean(birth_male),
            ratio = 100*share_male/(1 - share_male))
```

```
## # A tibble: 5 x 3
##   birth_order share_male ratio
##         <dbl>      <dbl> <dbl>
## 1           1      0.523  110.
## 2           2      0.514  106.
## 3           3      0.526  111.
## 4           4      0.525  111.
## 5           5      0.518  108.
```

No clear pattern $\rightarrow$ what's going on?

Sex ratio for first versus most recent birth

The birth order patterns mix mothers with different fertility plans

Might be better to compare first and most recent births

Generate dummy variable for first born

```
nfhs4 <-
  nfhs4 |>
  mutate(birth_first = if_else(birth_order==1, 1, 0))
```

Compare first and most recent births for mothers with at least 2 children

```
nfhs4 |>
  filter(mom_kids>=2 & (birth_order==1|birth_order==mom_kids)) |>
  group_by(birth_first) |>
  summarise(share_male = mean(birth_male),
            sex_ratio = 100*share_male/(1 - share_male))
```

```
## # A tibble: 2 x 3
##   birth_first share_male sex_ratio
##         <dbl>      <dbl>     <dbl>
## 1           0      0.583      140.
## 2           1      0.514      106.
```

The sex ratio at birth is much higher at the most recent birth

Sex ratio for by sex of older sibling

Generate dummy variable for the sex of the previous birth

```
nfhs4 <-
  nfhs4 |>
  group_by(mom_id) |>
  arrange(birth_order) |>
  mutate(older1_male = lag(birth_male)) |>
  ungroup()
```

Table by the sex of the the previous birth → higher ratio after girl

```
nfhs4 |>
  filter(birth_order>1) |>
  group_by(older1_male) |>
  summarise(share_male = mean(birth_male),
            sex_ratio = 100*share_male/(1 - share_male))
```

```
## # A tibble: 2 x 3
##   older1_male share_male sex_ratio
##         <dbl>      <dbl>     <dbl>
## 1           0      0.527      111.
## 2           1      0.511      104.
```

Son-biased fertility stopping

Parents want sons $\rightarrow$ more likely to continue childbearing after girl

Generate a dummy for whether there is a next birth

```
nfhs4 <-
  nfhs4 |>
  mutate(next_birth = if_else(mom_kids>birth_order, 1, 0))
```

Tabulate fertility continuation after a boy versus after a girl

Focus on births $\geq$ 2 years before survey, to allow time for next pregnancy

```
nfhs4 |>
  filter(birth_year<=2013) |>
  group_by(birth_male) |>
  summarise(share_continue = mean(next_birth))
```

```
## # A tibble: 2 x 2
##   birth_male share_continue
##        <dbl>          <dbl>
## 1          0          0.732
## 2          1          0.640
```

Sex of older siblings

Do these patterns intensify as parents try and try again for a boy?

To answer, let's create variables for the sexes of more older siblings

```
nfhs4 <-
  nfhs4 |>
  group_by(mom_id) |>
  arrange(birth_order) |>
  mutate(older2_male = lag(birth_male, n=2),
         older3_male = lag(birth_male, n=3),
         older4_male = lag(birth_male, n=4)
         ) |>
  ungroup()
```

Son-biased fertility stopping by birth order: table

Now let's create a new table

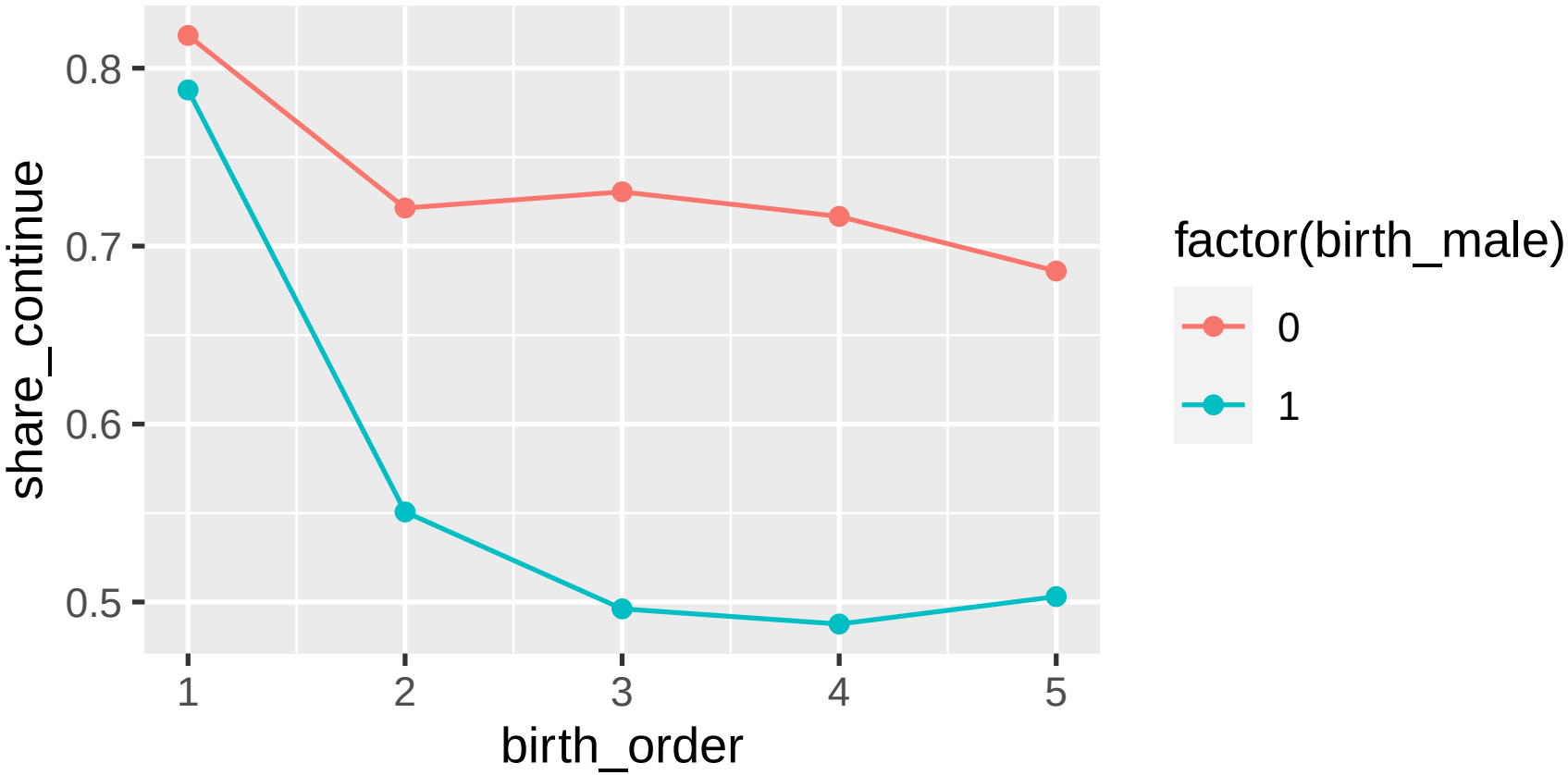
- fertility continuation after a boy or girl is born
- only for families that previously only had girls

```
table <-
 nfhs4 |>
 filter(birth_order==1 |
  birth_order==2&older1_male==0 |
  birth_order==3&older1_male==0&older2_male==0 |
  birth_order==4&older1_male==0&older2_male==0&older3_male==0 |
  birth_order==5&older1_male==0&older2_male==0&older3_male==0&older4_ma
 group_by(birth_order, birth_male) |>
 summarise(share_continue = mean(next_birth))
```

Son-biased fertility stopping by birth order: graph

Plot fertility continuation probabilities by birth order

```
ggplot(table, aes(x = birth_order, y = share_continue, color = factor(b
  geom_line() +
  geom_point()
```



Huge diffs in continuation for families that have had many girls, no boys

After 2 girls, birth of another girl raises Pr[another birth] by ~20 pp, or 40%

Sibship sizes for firstborn boys and girls: averages

Son-biased fertility stopping results in girls having larger families on average

For a very clean representation, let's look at...

- firstborn boys and girls
- mothers at the end of the reproductive period (45+)

```
nfhs4 |>
  filter(birth_order==1 & mom_age>=45) |>
  group_by(birth_male) |>
  summarise(mean_kids = mean(mom_kids))
```

```
## # A tibble: 2 x 2
##   birth_male mean_kids
##        <dbl>     <dbl>
## 1          0      3.90
## 2          1      3.43
```

Firstborn girls have 0.5 siblings more than firstborn boys

Fewer resources available per child

Sibship sizes for firstborn boys and girls: distributions

Instead of the means, let's plot histograms

Start by creating a table

- The piping here is more complicated than usual
- Count number of births with each `birth_male` by `mom_kids` combination
- Compute share for `birth_male==1` separately from `birth_male==0`

```
table <-
  nfhs4 |>
  filter(birth_order==1 & mom_age>=45) |>
  group_by(birth_male, mom_kids) |>
  summarise(n = n()) |>
  group_by(birth_male) |>
  mutate(share = n/sum(n))
```

Let's look at the table we just created

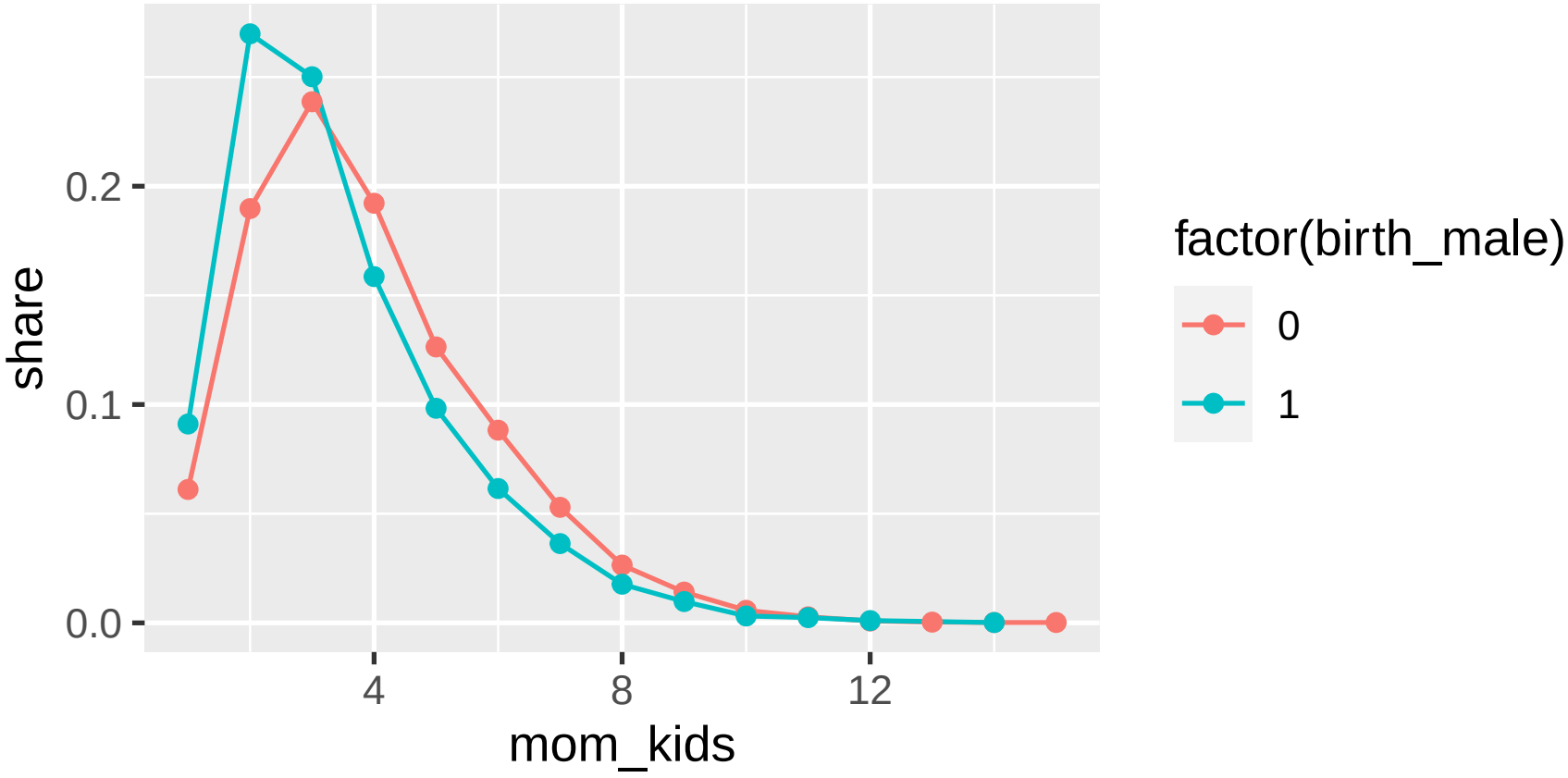
```
table
```

```
## # A tibble: 28 x 4
## # Groups:   birth_male [2]
##    birth_male mom_kids     n   share
##         <dbl>    <dbl> <int>   <dbl>
##  1          0        1   351 0.0611
##  2          0        2  1090 0.190
##  3          0        3  1371 0.239
##  4          0        4  1104 0.192
##  5          0        5   726 0.126
##  6          0        6   507 0.0883
##  7          0        7   304 0.0529
##  8          0        8   152 0.0265
##  9          0        9    81 0.0141
## 10          0       10    33 0.00575
## # i 18 more rows
```

Sibship sizes for firstborn boys and girls: distributions

Now plot the shares using `geom_line()` and `geom_point()`

```
ggplot(table, aes(x=mom_kids, y=share, color=factor(birth_male))) +
  geom_line() +
  geom_point()
```



Can see how the distribution of sibship size is shifted to the right for girls

Sibship sizes for all boys and girls: averages

Now let's check the patterns for **all** boys and girls

```
nfhs4 |>
  group_by(birth_male) |>
  summarise(mean_kids = mean(mom_kids))
```

```
## # A tibble: 2 x 2
##   birth_male mean_kids
##        <dbl>     <dbl>
## 1          0      3.78
## 2          1      3.49
```

Still true that girls have larger families than boys on average

Difference is somewhat smaller now, but many families are not complete

Sibship size and infant mortality

We saw that girls tend to have larger families than boys

I suggested that this pattern adds to their disadvantage through resource dilution

Relationship between sibship size and infant mortality?

```
nfhs4 |>
  group_by(mom_kids) |>
  summarise(mort_rate = mean(birth_u1)) |>
  filter(mom_kids<=5)
```

```
## # A tibble: 5 x 2
##   mom_kids mort_rate
##      <dbl>     <dbl>
## 1        1    0.0231
## 2        2    0.0176
## 3        3    0.0383
## 4        4    0.0615
## 5        5    0.0793
```

Infant mortality much more common in large families!

But is this a causal effect of sibship size? Hint: probably not. . .

Sibship size, birth order, and infant mortality

Birth order is intrinsically related to sibship size

Can't be 10^{th} born without being in a large family

Consider the following table:

Sibship size	Possible birth orders	Average birth order
1	1	1
2	1, 2	1.5
3	1, 2, 3	2.
4	1, 2, 3, 4	2.5
5	1, 2, 3, 4, 5	3
⋮	⋮	⋮
10	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	5.5

Or...

$$\overline{order} = \frac{1}{2}\left(1 + sibsize\right)$$

implying that sibship size differences may in part reflect birth order differences

Sibship size, birth order, and infant mortality: table

Let's try to distinguish birth order from sibship size

Note: this exercise is a tangent! We'll get back to studying gender bias soon

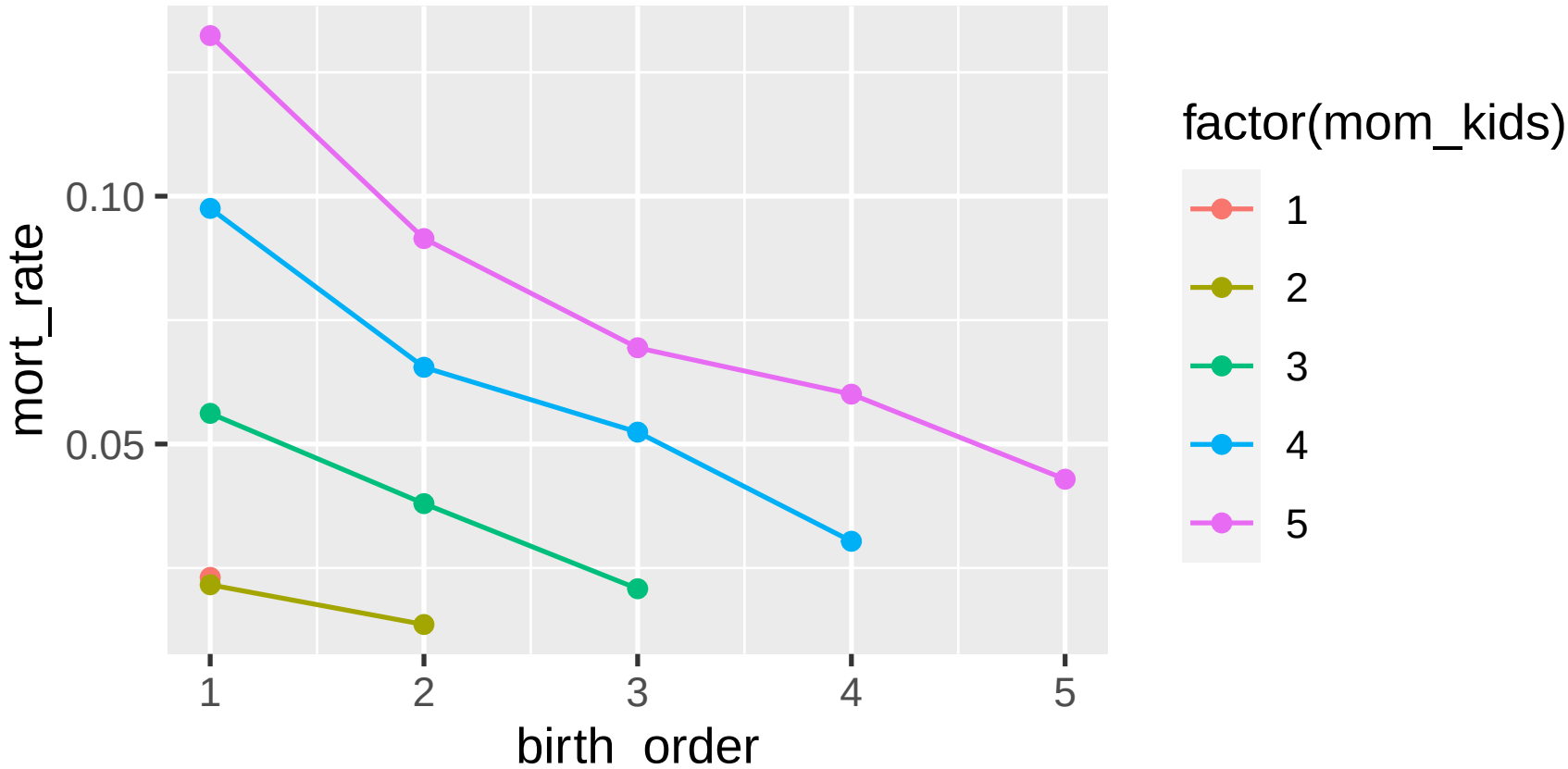
Let's make a table of mortality risk by sibship size and birth order

```
table <-
  nfhs4 |>
  group_by(mom_kids, birth_order) |>
  summarise(mort_rate = mean(birth_u1)) |>
  filter(mom_kids<=5)
```

Sibship size, birth order, and infant mortality: graph

Now draw a graph

```
ggplot(table, aes(x=birth_order, y=mort_rate, color=factor(mom_kids)))
  geom_line() +
  geom_point()
```



Within a given family size, later-born are **less** likely to die

$\rightarrow$ at a given birth order, family size differences are even **larger**

Sex differences in infant mortality

Does son preference lead to higher mortality for girls?

```
nfhs4 |>
  group_by(birth_male) |>
  summarise(mortality_rate = mean(birth_u1))
```

```
## # A tibble: 2 x 2
##   birth_male mortality_rate
##        <dbl>          <dbl>
## 1          0         0.0461
## 2          1         0.0552
```

No! Remember that infant boys are naturally more frail than infant girls

Sibship size, birth order, sex, and infant mortality: table

Let's recreate the most recent graph with separate panels for boys and girls

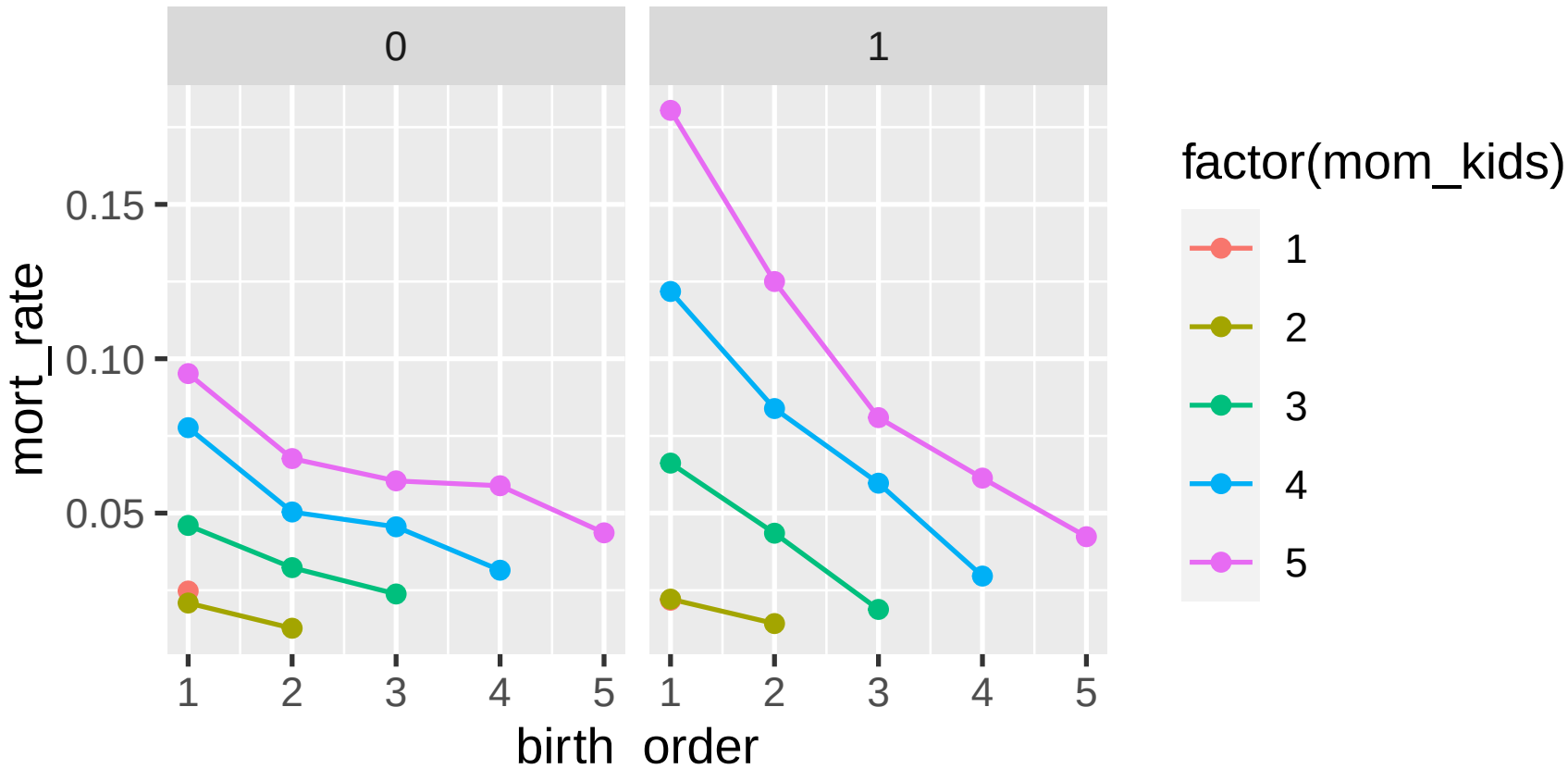
Same code as before, just adding `birth_male` to the `group_by()` line

```
table <-
  nfhs4 |>
  group_by(mom_kids, birth_order, birth_male) |>
  summarise(mort_rate = mean(birth_u1)) |>
  filter(mom_kids<=5 & birth_order<=5)
```

Sibship size, birth order, sex, and infant mortality: graph

Now let's use `table` to draw separate graphs for boys and girls

```
ggplot(table, aes(x=birth_order, y=mort_rate, color=factor(mom_kids)))
  geom_line() +
  geom_point() +
  facet_wrap(~birth_male)
```



Larger diffs in boys' risk across family sizes

Sex differences in infant mortality following a boy or girl

Let's dig deeper → sex diffs in infant mortality by sex of older sibling

```
nfhs4 |>
  filter(birth_order>=2) |>
  group_by(older1_male, birth_male) |>
  summarise(mortality_rate = mean(birth_u1))
```

```
## # A tibble: 4 x 3
## # Groups:   older1_male [2]
##   older1_male birth_male mortality_rate
##         <dbl>      <dbl>          <dbl>
## 1           0          0         0.0461
## 2           0          1         0.0468
## 3           1          0         0.0448
## 4           1          1         0.0579
```

No mortality gap the older sibling is a sister → excess girl mortality

Girls with older sisters are disadvantaged

- Parents who continue after a daughter are more son-biased on average
- Parents may under-invest in daughters' health while they try for a son

Sex differences in infant mortality by birth order: table

New table by birth order

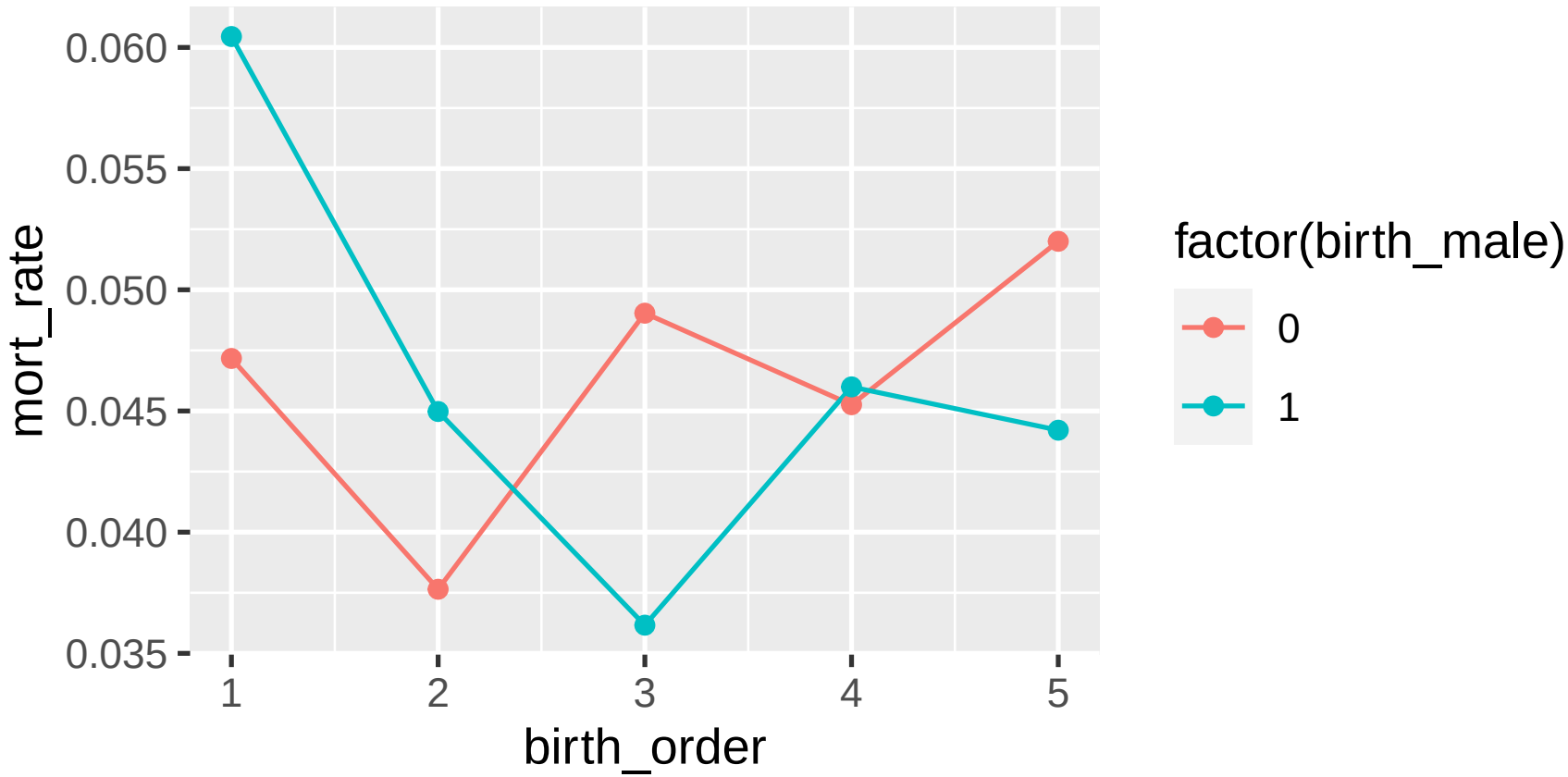
- infant mortality for boys versus girls
- only for families that previously only had girls

```
table <-
 nfhs4 |>
 filter(birth_order==1 |
  birth_order==2&older1_male==0 |
  birth_order==3&older1_male==0&older2_male==0 |
  birth_order==4&older1_male==0&older2_male==0&older3_male==0 |
  birth_order==5&older1_male==0&older2_male==0&older3_male==0&older4_ma
 group_by(birth_order, birth_male) |>
 summarise(mort_rate = mean(birth_u1))
```

Sex differences in infant mortality by birth order: graph

If many older sisters and no older brothers, girls are more likely to die than boys

```
ggplot(table, aes(x = birth_order, y = mort_rate, color = factor(birth_
  geom_line() +
  geom_point()
```



Concluding thoughts

Many patterns consistent with son-preference:

- Lastborns more likely to be boys than firstborns
- After the birth of a girl (rather than a boy)...
  - Parents are more likely to have more another child
  - Next-born more likely to be a boy
  - Girl-advantage in infant mortality erased for next-born
- These patterns intensify after the birth of many girls