SOC-5811 Week 3: Quantitative research methods

Nick Graetz

University of Minnesota, Department of Sociology

9/17/2025

Quantitative methods

- **Empiricism:** social theories imply empirical claims that are falsifiable.

Quantitative methods

- **Empiricism:** social theories imply empirical claims that are falsifiable.
- Data doesn't speak for itself; it must be carefully interpreted, summarize, and analyzed.

Quantitative methods

- Statistics is about providing a concise summary of empirical data.

Quantitative methods

- Statistics is about providing a concise summary of empirical data.
- This involves analyzing **samples** drawn from **populations**.

Samples and populations

Say I have a bag filled with 50 marbles of different colors. I reach in the bag and pull out 10 marbles. I see that 6 marbles are blue and 4 are red.

- What is the population and sample here?

Samples and populations

Say I have a bag filled with 50 marbles of different colors. I reach in the bag and pull out 10 marbles. I see that 6 marbles are blue and 4 are red.

- What is the population and sample here?
- What can I say about the sample? What can I say about the population?

Generalization

Summarizing in-sample

- **Measurement**: generalizing from **observed measurements** to the **underlying constructs of interest**.

Generalization

Summarizing in-sample

- **Measurement**: generalizing from **observed measurements** to the **underlying constructs of interest**.
  - Is what I'm measuring what I actually care about?

Generalization

Summarizing in-sample

- **Measurement**: generalizing from **observed measurements** to the **underlying constructs of interest**.
  - Is what I'm measuring what I actually care about?
  - *I ask 30 people whether they graduated from college; what am I measuring?*

Generalization

Summarizing in-sample

- **Measurement**: generalizing from **observed measurements** to the **underlying constructs of interest**.
  - Is what I'm measuring what I actually care about?
  - *I ask 30 people whether they graduated from college; what am I measuring?*
  - Measurement theory: very central to psychology, education, etc.

Generalization

Summarizing in-sample

- **Measurement**: generalizing from **observed measurements** to the **underlying constructs of interest**.
  - Is what I'm measuring what I actually care about?
  - *I ask 30 people whether they graduated from college; what am I measuring?*
  - Measurement theory: very central to psychology, education, etc.

- **Generalizing from sample to population**.

Generalization

Summarizing in-sample

- **Measurement**: generalizing from **observed measurements** to the **underlying constructs of interest**.
  - Is what I'm measuring what I actually care about?
  - *I ask 30 people whether they graduated from college; what am I measuring?*
  - Measurement theory: very central to psychology, education, etc.

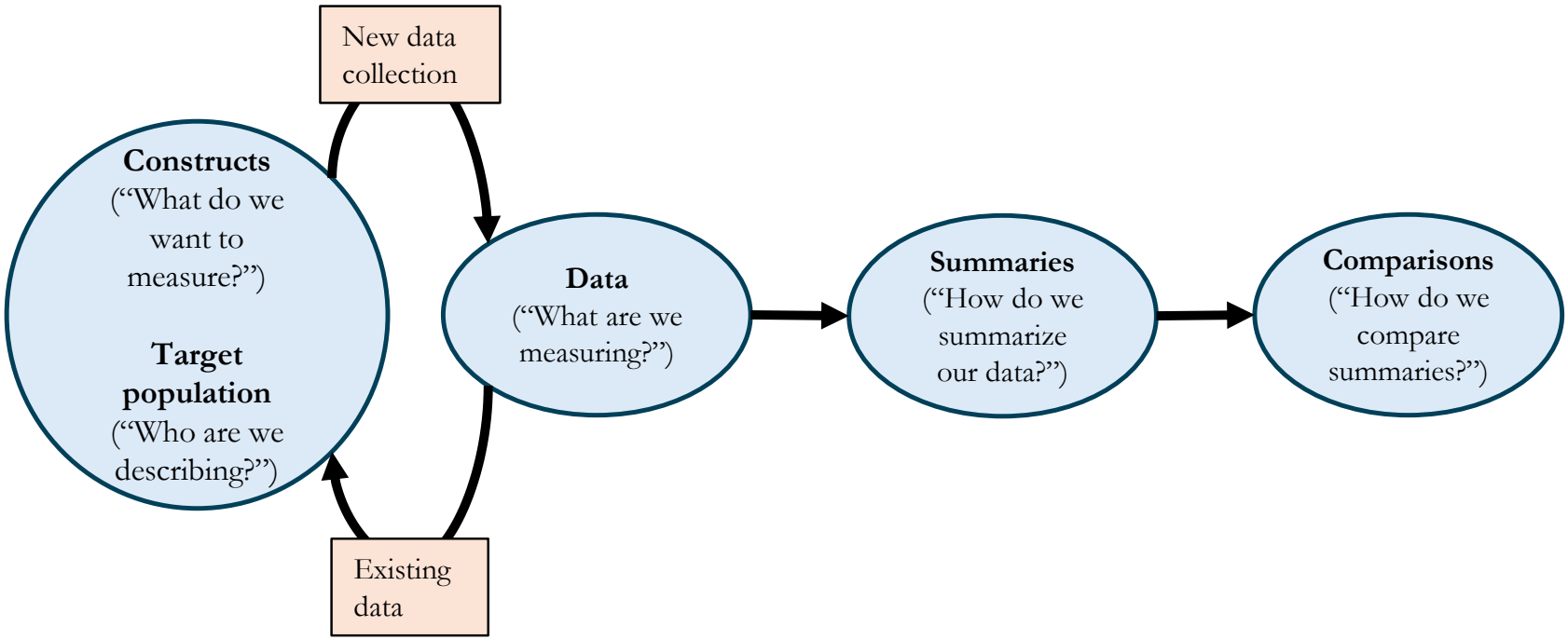
- **Generalizing from sample to population**.
  - Is what I see in the sample also true in the population?

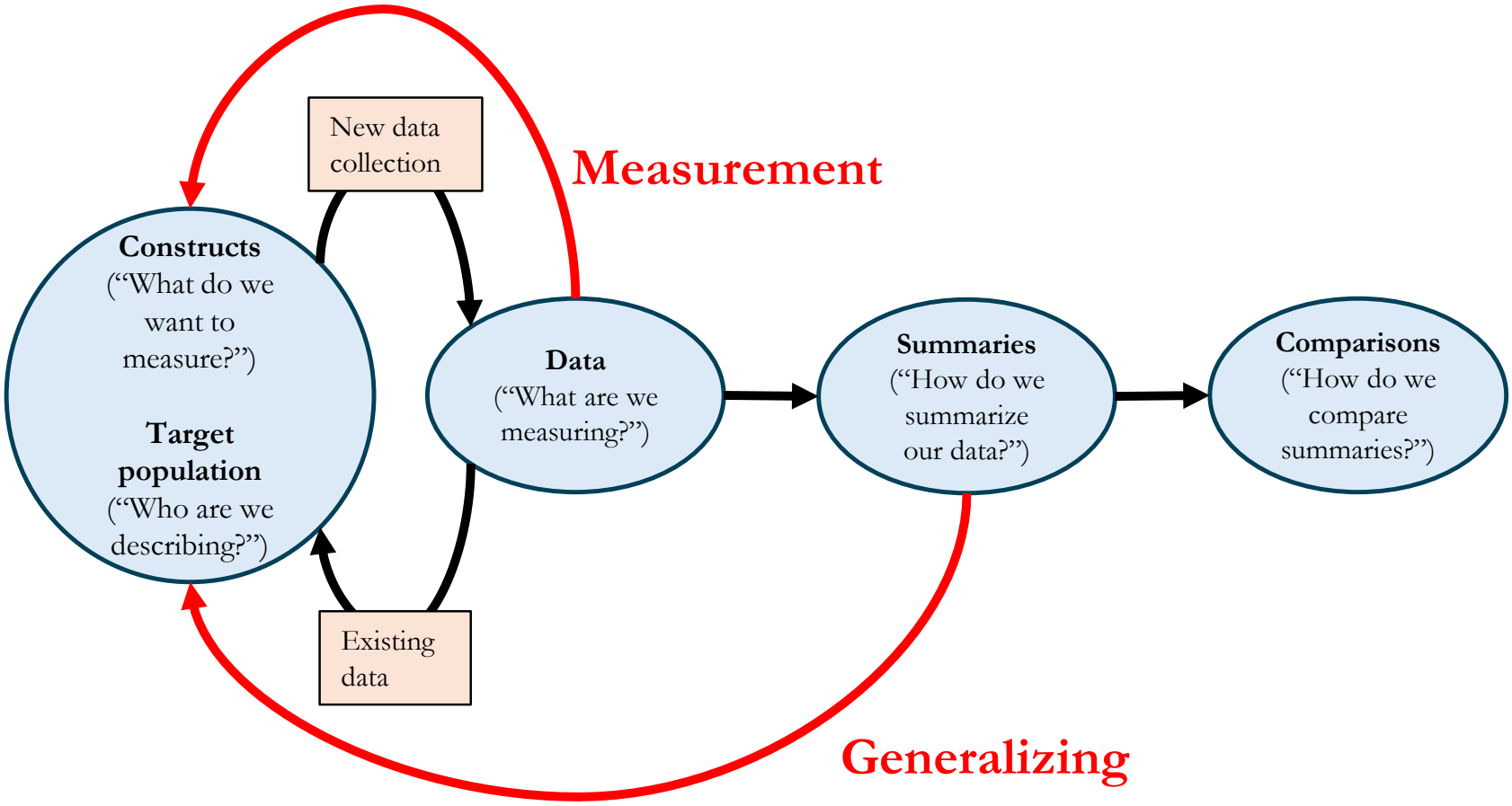
Generalization

Summarizing in-sample

- **Measurement**: generalizing from **observed measurements** to the **underlying constructs of interest**.
  - Is what I'm measuring what I actually care about?
  - *I ask 30 people whether they graduated from college; what am I measuring?*
  - Measurement theory: very central to psychology, education, etc.

- **Generalizing from sample to population**.
  - Is what I see in the sample also true in the population?
  - *I ask 30 people whether they graduated from college; what can I say about the population from which I sampled these people?*


New data collection
Constructs ("What do we want to measure?")
Target population ("Who are we describing?")
Data ("What are we measuring?")
Summaries ("How do we summarize our data?")
Comparisons ("How do we compare summaries?")
Existing data


New data collection
Measurement
Constructs
("What do we want to measure?")
Target population
("Who are we describing?")
Data
("What are we measuring?")
Summaries
("How do we summarize our data?")
Comparisons
("How do we compare summaries?")
Existing data
Generalizing

Generalization

- **Forecasting**: generalizing from the present to the future.

Generalization

- **Forecasting**: generalizing from the present to the future.
  - *I have data on whether on not it has rained every day for the past ten years. Will it rain tomorrow?*

Generalization

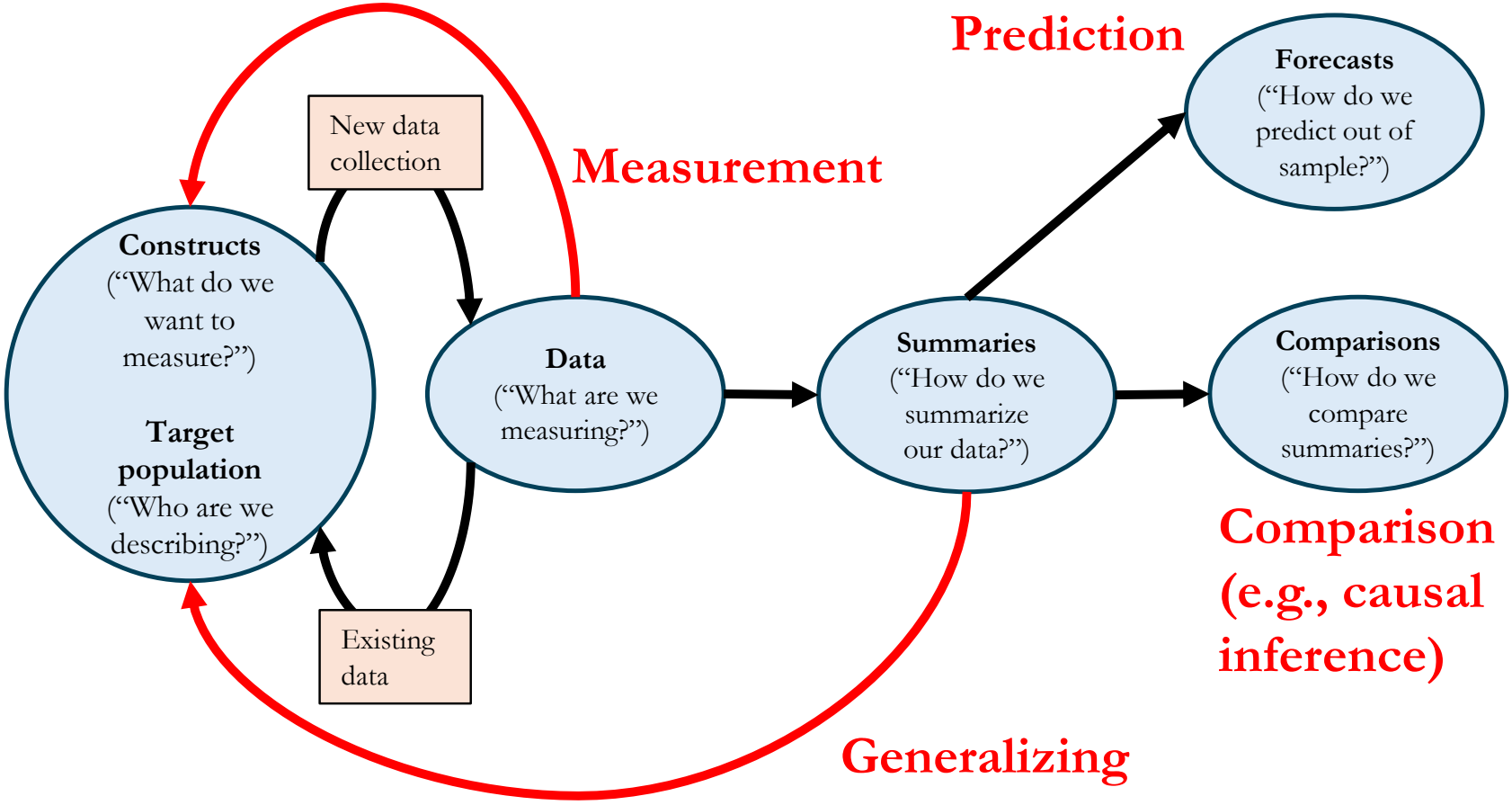
- **Forecasting**: generalizing from the present to the future.
  - *I have data on whether on not it has rained every day for the past ten years. Will it rain tomorrow?*

- **Causal inference**: generalizing from treatment group to control group.

Generalization

- **Forecasting**: generalizing from the present to the future.
  - *I have data on whether on not it has rained every day for the past ten years. Will it rain tomorrow?*

- **Causal inference**: generalizing from treatment group to control group.
  - *I measure the rate of a specific disease among a group that was vaccinated for that disease and one that was not. Was the vaccine effective at reducing the disease?*


Prediction
Forecasts
("How do we predict out of sample?")
New data collection
Measurement
Constructs
("What do we want to measure?")
Target population
("Who are we describing?")
Data
("What are we measuring?")
Summaries
("How do we summarize our data?")
Comparisons
("How do we compare summaries?")
Comparison (e.g., causal inference)
Existing data
Generalizing

Generalization: Summary

1. Measurement.
2. Generalizing from sample to population.
3. Comparison (e.g., causal inference).
4. Forecasting.

Generalization: Summary

1. Measurement.
2. **Generalizing from sample to population.**
3. **Comparison (e.g., causal inference.)**
4. Forecasting.

Quantitative data

Data on US populations by state/region

```
census <- read_dta(paste0(data_dir,
                          's5811-msa.dta')) %>%
  mutate(across(c(state,region,isusa), as_label)) %>%
  head()
```

Quantitative data

Continuous and discrete variables

```
census %>%
  select(population,region) %>%
  slice(1:5)
```

```
## # A tibble: 5 x 2
##   population region
##        <dbl> <fct>
## 1      5949. 1_Northeast
## 2      1919. 4_West
## 3      5545. 3_South
## 4      2271. 3_South
## 5       724. 3_South
```

Levels of measurement

- **Continuous variables:** infinite number of values in a given interval (e.g., time)

Levels of measurement

- **Continuous variables:** infinite number of values in a given interval (e.g., time)
- **Discrete variables:** limited to a specific set of outcomes (e.g., die roll)

Levels of measurement

- **Continuous variables:** infinite number of values in a given interval (e.g., time)
- **Discrete variables:** limited to a specific set of outcomes (e.g., die roll)
  - Ordinal: ordered, discrete categories (e.g., ranks in a race)

Levels of measurement

- **Continuous variables:** infinite number of values in a given interval (e.g., time)
- **Discrete variables:** limited to a specific set of outcomes (e.g., die roll)
  - Ordinal: ordered, discrete categories (e.g., ranks in a race)
  - Nominal: unordered, discrete categories (e.g., field of study)

Levels of measurement

- **Continuous variables:** infinite number of values in a given interval (e.g., time)
- **Discrete variables:** limited to a specific set of outcomes (e.g., die roll)
  - Ordinal: ordered, discrete categories (e.g., ranks in a race)
  - Nominal: unordered, discrete categories (e.g., field of study)
  - Binary/dichotomous: special case of nominal variable that only has two values (e.g., coin flip)

Levels of measurement

- **Continuous variables:** infinite number of values in a given interval (e.g., time)
- **Discrete variables:** limited to a specific set of outcomes (e.g., die roll)
  - Ordinal: ordered, discrete categories (e.g., ranks in a race)
  - Nominal: unordered, discrete categories (e.g., field of study)
  - Binary/dichotomous: special case of nominal variable that only has two values (e.g., coin flip)
- Measurement decisions are not always clear cut (e.g., age)

SUMMARIZING DATA

```
head(gss)
```

```
## # A tibble: 6 x 8
##   degree       agekdbrn  realrinc  race  age
##   <fct>        <dbl+lbl> <dbl+lbl> <fct> <dbl+lbl>
## 1 bachelor     35        45400     white 42
## 2 bachelor     32        54480     white 63
## 3 lt high school 17        908     white 62
## 4 high school  30        45400     white 55
## 5 graduate     30        54480     white 59
## 6 high school  20         8512.    other 34
```

Summarizing data

```
gss %>%
  summarise(mean_income=mean(realrinc))
```

```
## # A tibble: 1 x 1
##   mean_income
##         <dbl>
## 1          NA
```

Summarizing data

```
gss %>%
  summarise(mean_income=mean(realrinc, na.rm=T))
```

```
## # A tibble: 1 x 1
##   mean_income
##         <dbl>
## 1       25750.
```

SUMMARIZING DATA

```
gss %>%
  group_by(degree) %>%
  summarise(mean_income=mean(realrinc, na.rm=T))
```

```
## # A tibble: 5 x 2
##   degree          mean_income
##   <fct>                 <dbl>
## 1 lt high school       12888.
## 2 high school          18795.
## 3 junior college       18636.
## 4 bachelor             36143.
## 5 graduate             46767.
```

SUMMARIZING DATA

```
gss %>%
  group_by(degree) %>%
  summarise(n=n()) %>%
  mutate(prop=n/sum(n)*100)
```

```
## # A tibble: 5 x 3
##   degree              n  prop
##   <fct>           <int> <dbl>
## 1 lt high school    153 10.7
## 2 high school       724 50.8
## 3 junior college    117  8.20
## 4 bachelor          282 19.8
## 5 graduate          150 10.5
```

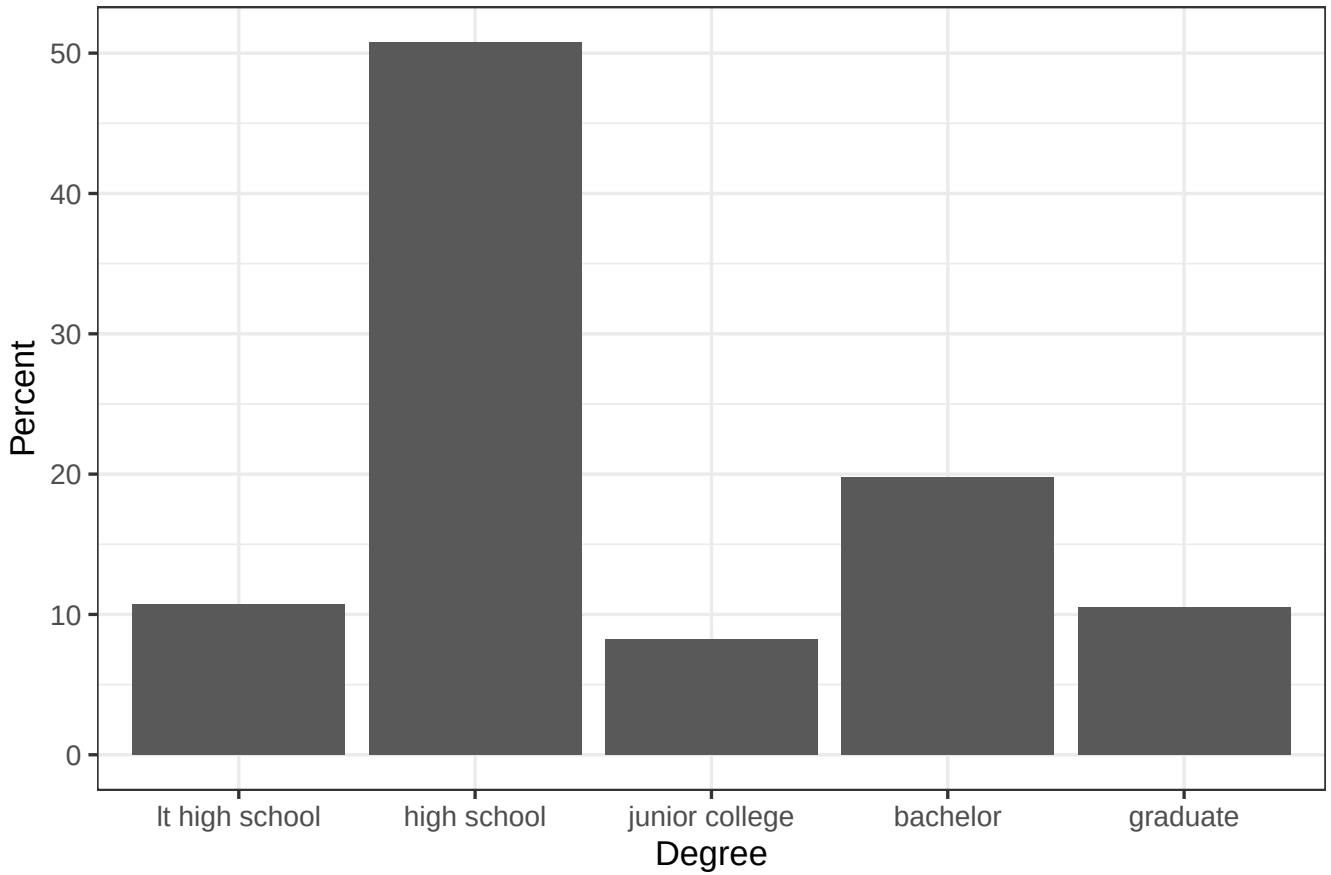
Summarizing data

```
plot_data <- gss %>%
  group_by(degree) %>%
  summarise(n=n()) %>%
  mutate(prop=n/sum(n)*100)
```

Summarizing data

```
plot <- ggplot(data=plot_data,
                  aes(x=degree,y=prop)) +
  geom_bar(stat='identity') +
  labs(x='Degree',y='Percent') +
  theme_bw()
```

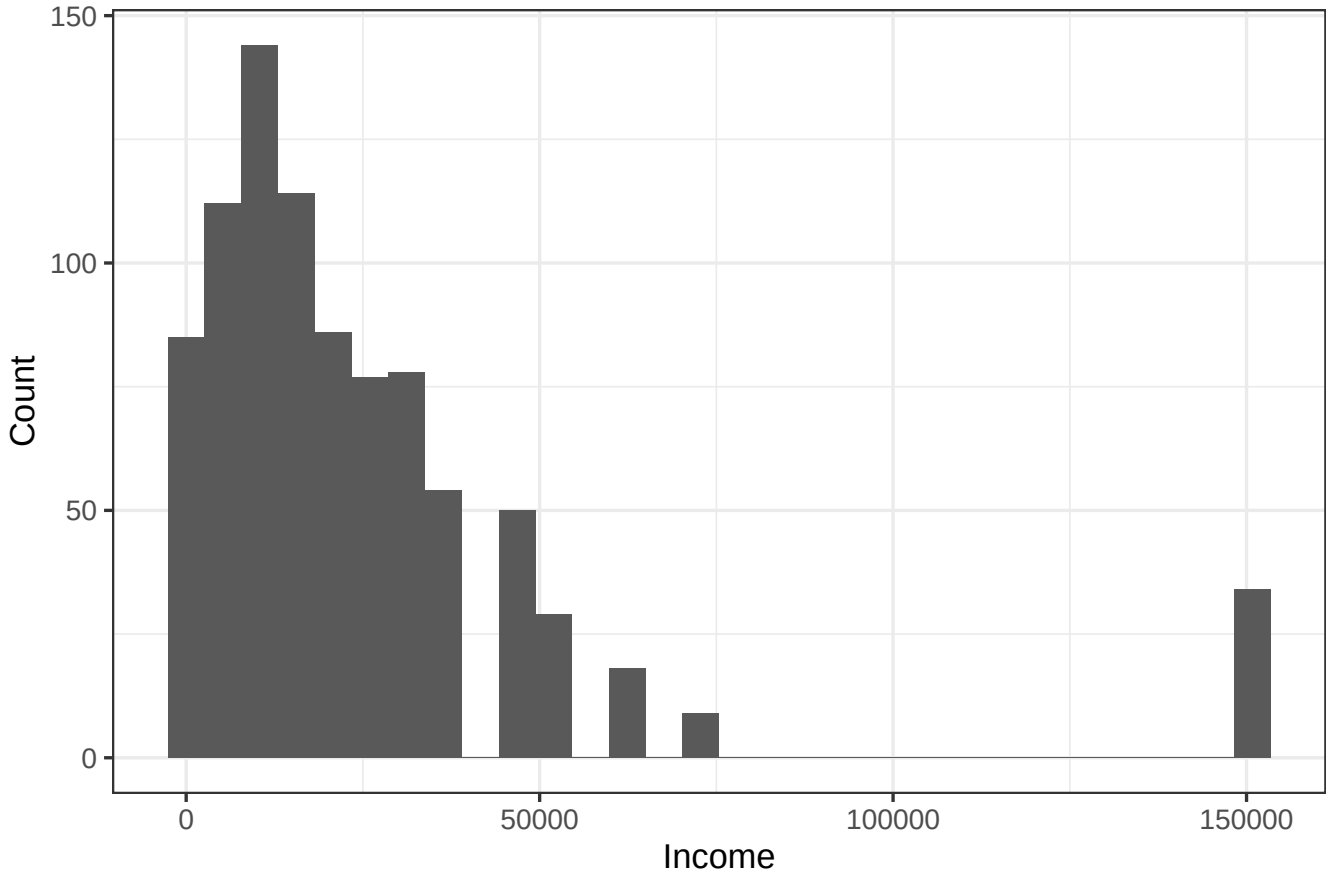
Summarizing data



Summarizing data

```
plot <- ggplot(data=gss,
                  aes(x=realrinc)) +
  geom_histogram() +
  labs(x='Income',y='Count') +
  theme_bw()
```

Summarizing data



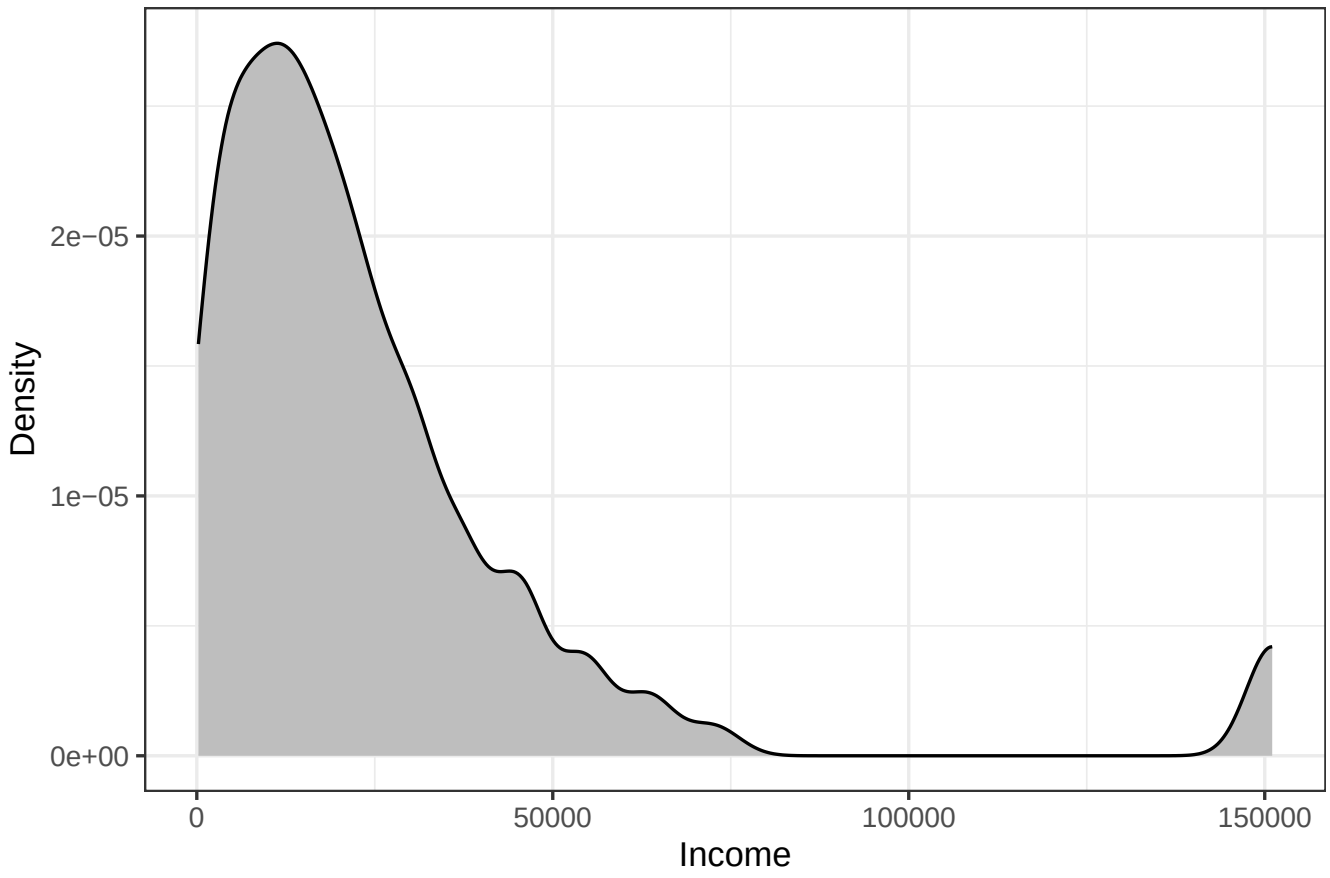
Summarizing data

- Histograms of continuous data: why do we have to bin?
- How do we choose bins?

Summarizing data

```
plot <- ggplot(data=gss,
                  aes(x=realrinc)) +
  geom_density(fill='grey') +
  labs(x='Income',y='Density') +
  theme_bw()
```

Summarizing data



SUMMARIZING DATA

Add measures of central tendency: mean and median

```
summaries <- gss %>%
  summarise(mean_income=mean(realrinc, na.rm=T),
            median_income=median(realrinc, na.rm=T))
summaries
```

```
## # A tibble: 1 x 2
##   mean_income median_income
##         <dbl>         <dbl>
## 1      25750.         17025
```

Summarizing data

Reshape:

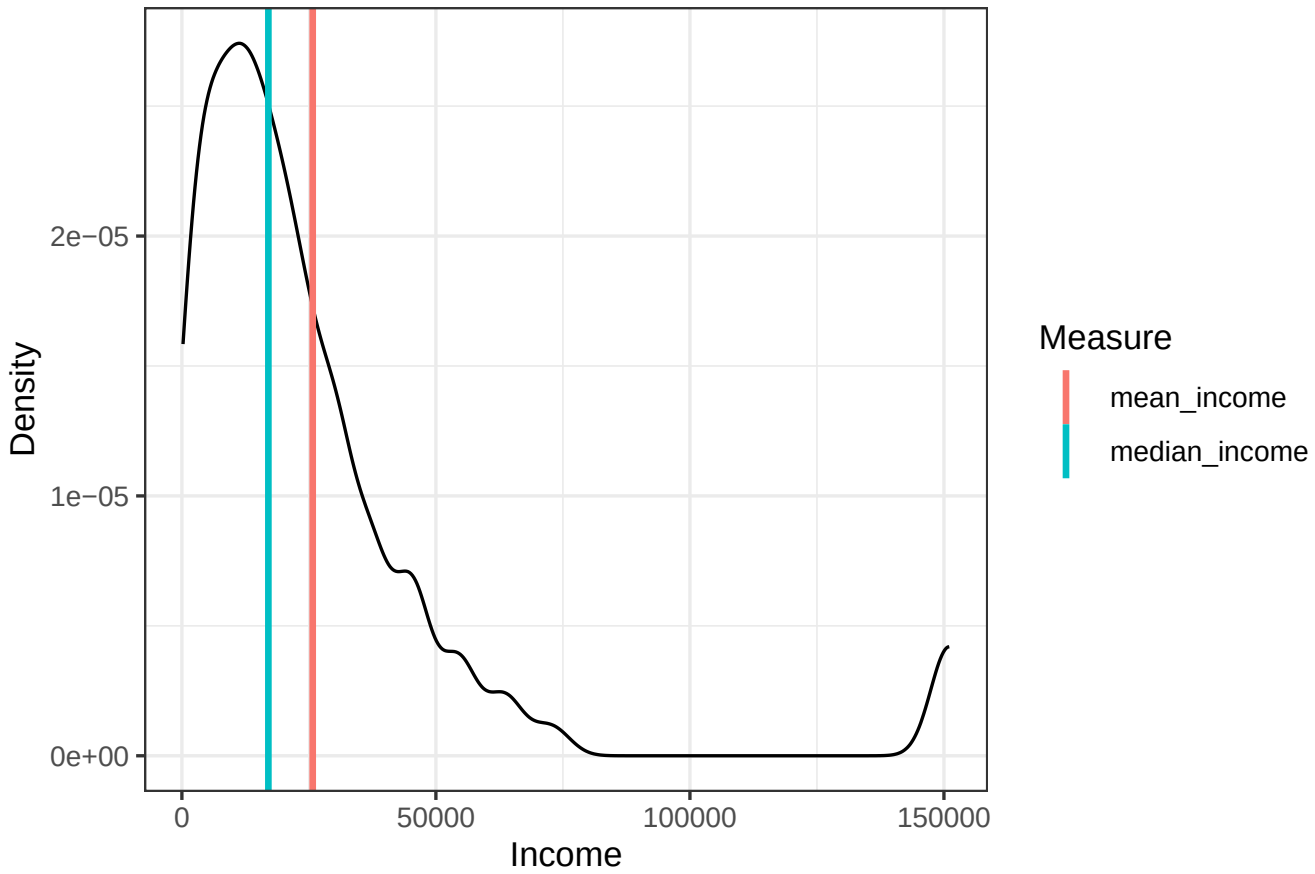
```
summaries <- gss %>%
  summarise(mean_income=mean(realrinc, na.rm=T),
            median_income=median(realrinc, na.rm=T)) %>%
  pivot_longer(cols=contains('income'),
               names_to='Measure',
               values_to='income')
summaries
```

```
## # A tibble: 2 x 2
##   Measure       income
##   <chr>          <dbl>
## 1 mean_income   25750.
## 2 median_income 17025
```

Summarizing data

```
plot <- ggplot(data=gss,
                  aes(x=realrinc)) +
  geom_density() +
  geom_vline(data=summaries,
             aes(xintercept=income,
                 color=Measure),
             size=1) +
  labs(x='Income',y='Density') +
  theme_bw()
```

Summarizing data



Review

What do we want to do with data?

- Manipulate variables, which might involve conditional statements

Review

What do we want to do with data?

- Manipulate variables, which might involve conditional statements
- **Merge:** combine two datasets on matching columns

Review

What do we want to do with data?

- Manipulate variables, which might involve conditional statements
- **Merge:** combine two datasets on matching columns
- **Reshape:** wide to long, long to wide

Review

What do we want to do with data?

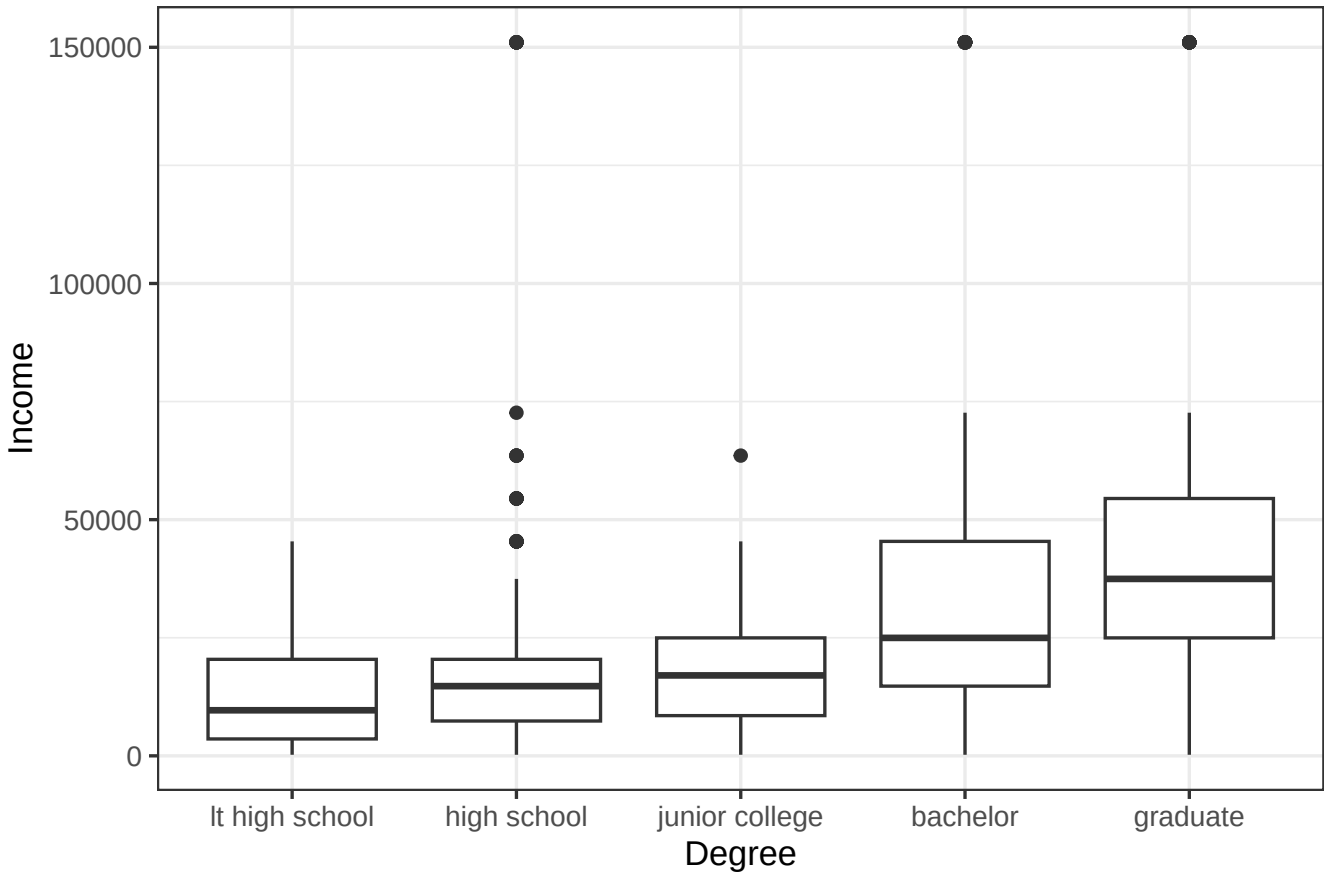
- Manipulate variables, which might involve conditional statements
- **Merge:** combine two datasets on matching columns
- **Reshape:** wide to long, long to wide
- **Summarize:** aggregate over rows

Univariate summary statistics

Examine distribution:

```
plot <- ggplot(data=gss,
                aes(y=realrinc,x=degree)) +
  geom_boxplot() +
  labs(x='Degree',y='Income') +
  theme_bw()
```

Univariate summary statistics



Bivariate relationships

Correlation matrix:

```
gss %>%
  select(educ,prestg10,realrinc,agekdbrn) %>%
  cor(use = 'complete.obs') %>%
  round(2)
```

```
##             educ prestg10 realrinc agekdbrn
## educ        1.00     0.49     0.35     0.42
## prestg10    0.49     1.00     0.30     0.37
## realrinc    0.35     0.30     1.00     0.22
## agekdbrn    0.42     0.37     0.22     1.00
```

BIVARIATE RELATIONSHIPS

Cross-tabulations table:

```
gss %>%
  select(degree,race) %>%
  table() %>%
  prop.table() %>%
  round(2)
```

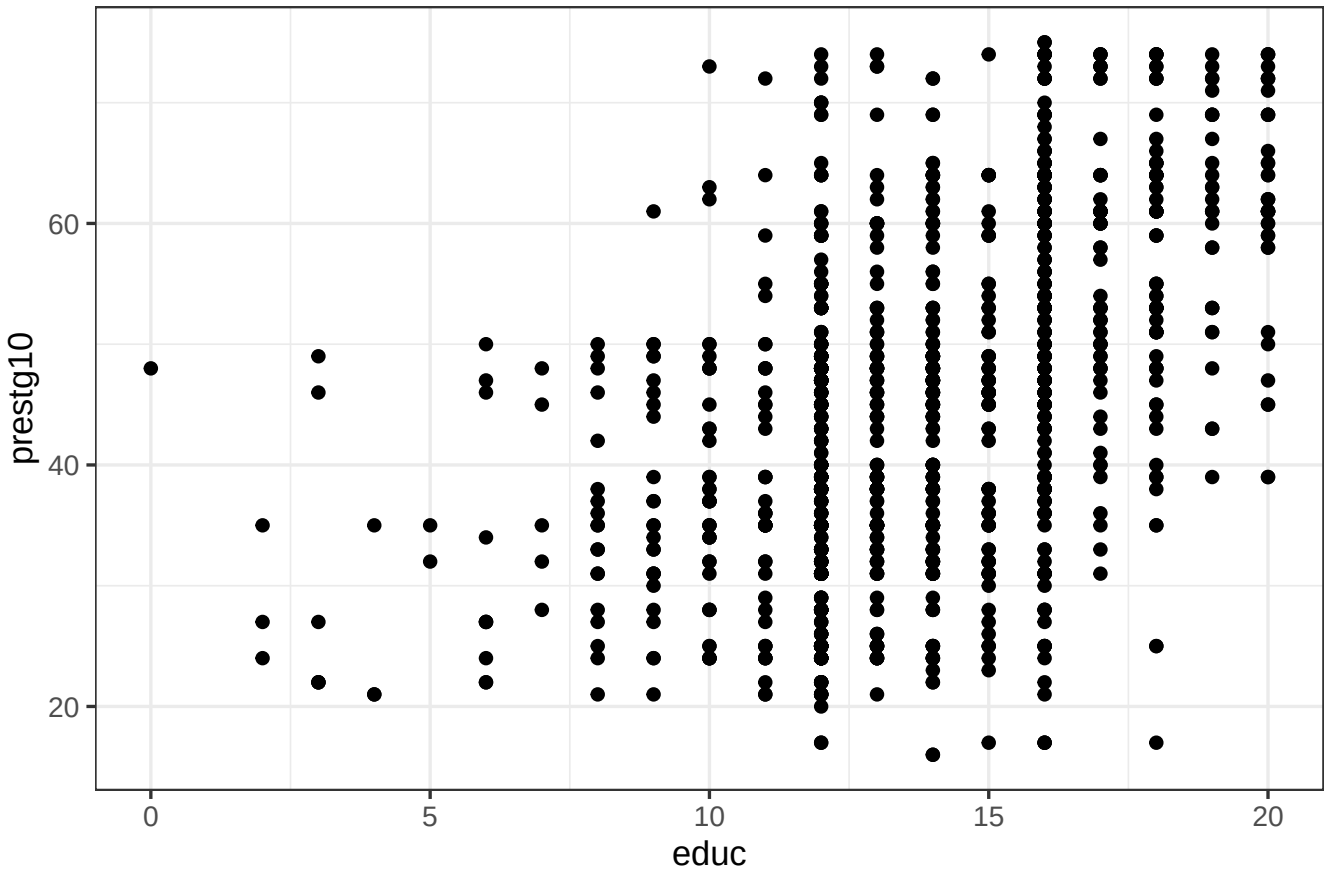
```
##                  race
## degree            white black other
##   lt high school   0.06  0.02  0.03
##   high school      0.36  0.09  0.05
##   junior college   0.06  0.01  0.00
##   bachelor         0.15  0.03  0.02
##   graduate         0.08  0.01  0.01
```

Bivariate relationships

Examine bivariate relationship:

```
plot <- ggplot(data=gss,
                aes(x=educ,y=prestg10)) +
  geom_point() +
  theme_bw()
```

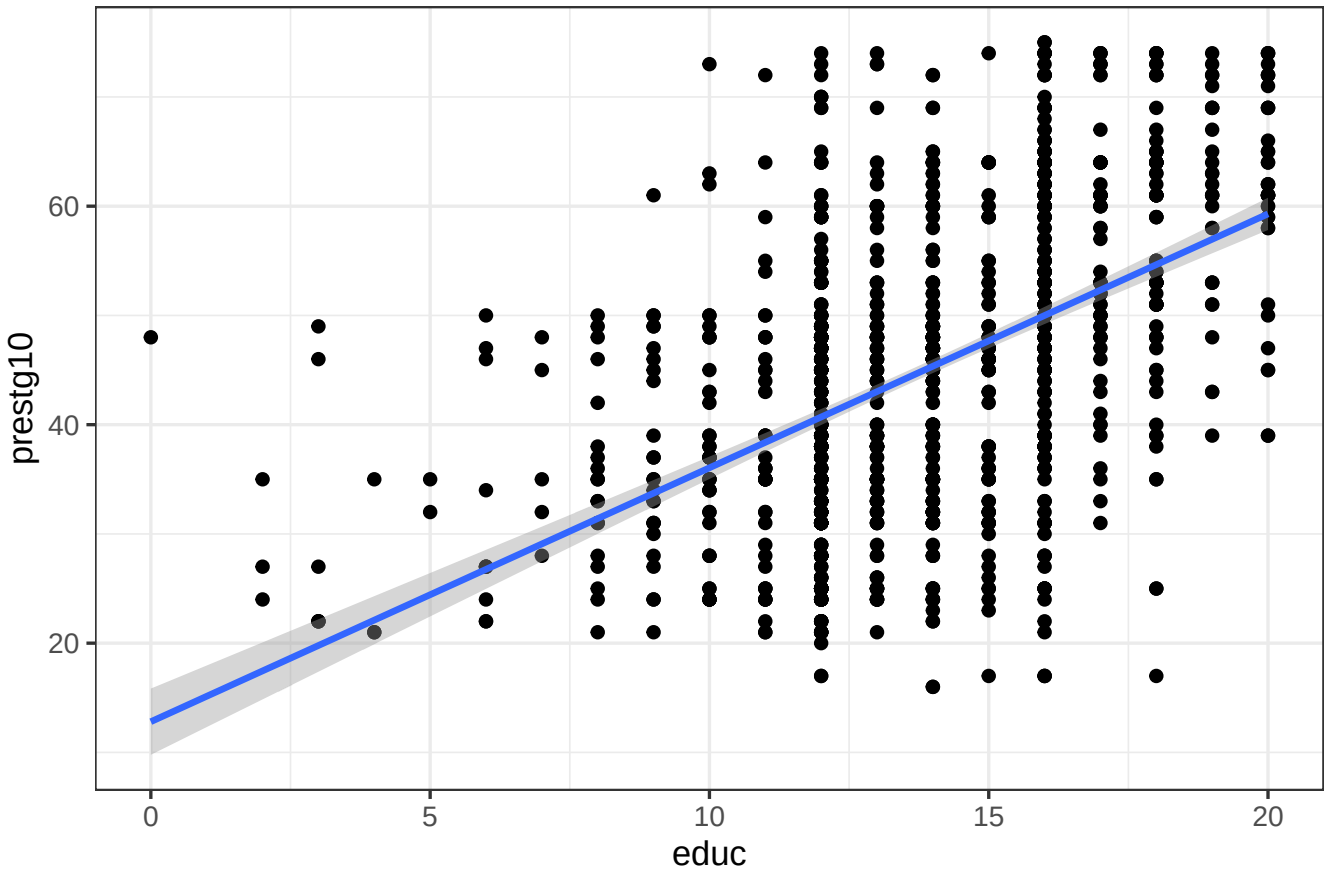
Bivariate relationships



Bivariate relationships

```
plot <- ggplot(data=gss,
                aes(x=educ,y=prestg10)) +
  geom_point() +
  geom_smooth(method='lm') +
  theme_bw()
```

Bivariate relationships



Stats and social science theory

How do we create models?

- Deterministic processes
- Stochastic processes