# Descriptive Statistics, pt. III

## ECON 3640–001

Marcio Santetti

Spring 2022

# Motivation

# The road so far

So far, our descriptive measures (e.g., *mean, median, variance, standard deviation*) suit well our purposes when describing a **unique** variable.

These measures are also known as **univariate** descriptive techniques.

Whenever our goal is to describe a possible *relationship/association* between two variables, we need to study additional descriptive techniques.

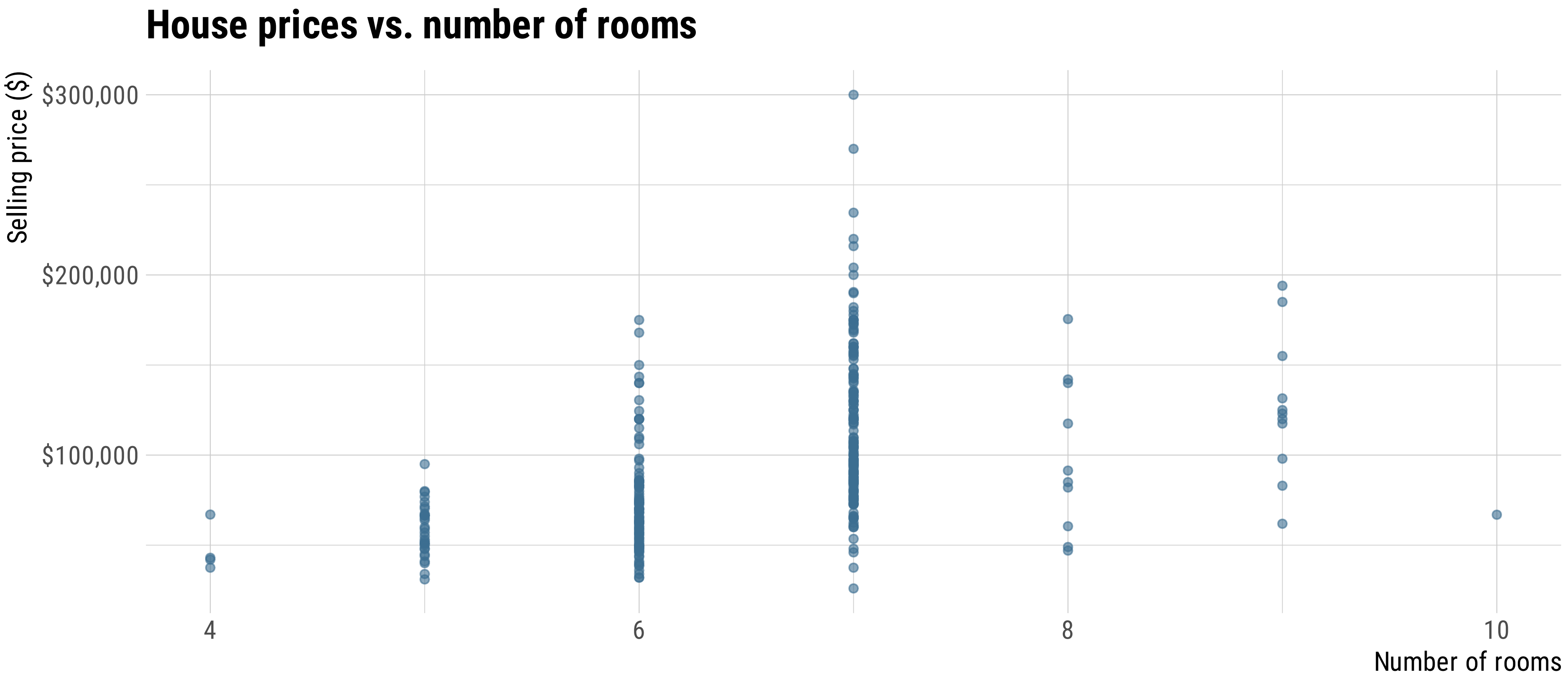
These are known as **bivariate** descriptive measures.

We will study the three main techniques:

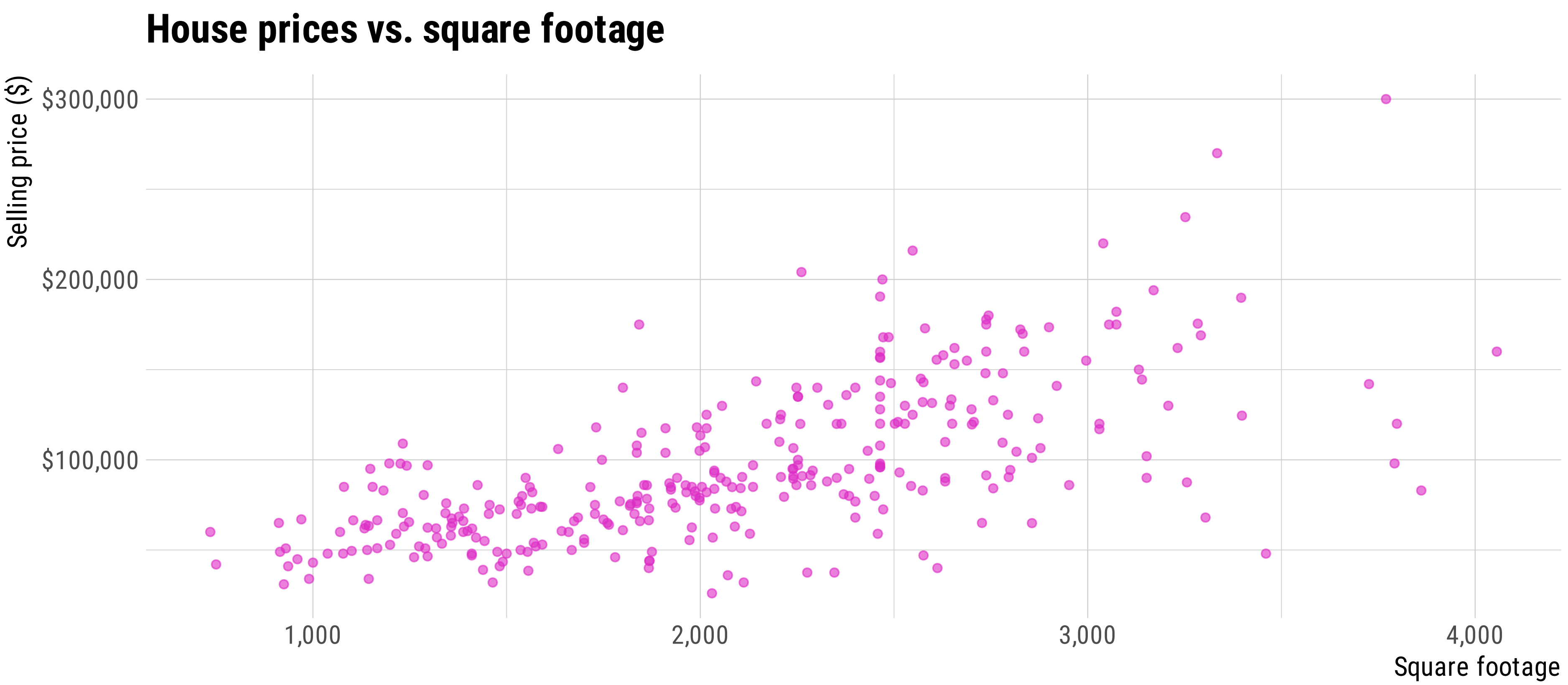
- *Covariance*;
- *Correlation*;
- The *coefficient of determination*.

# Bivariate descriptive techniques

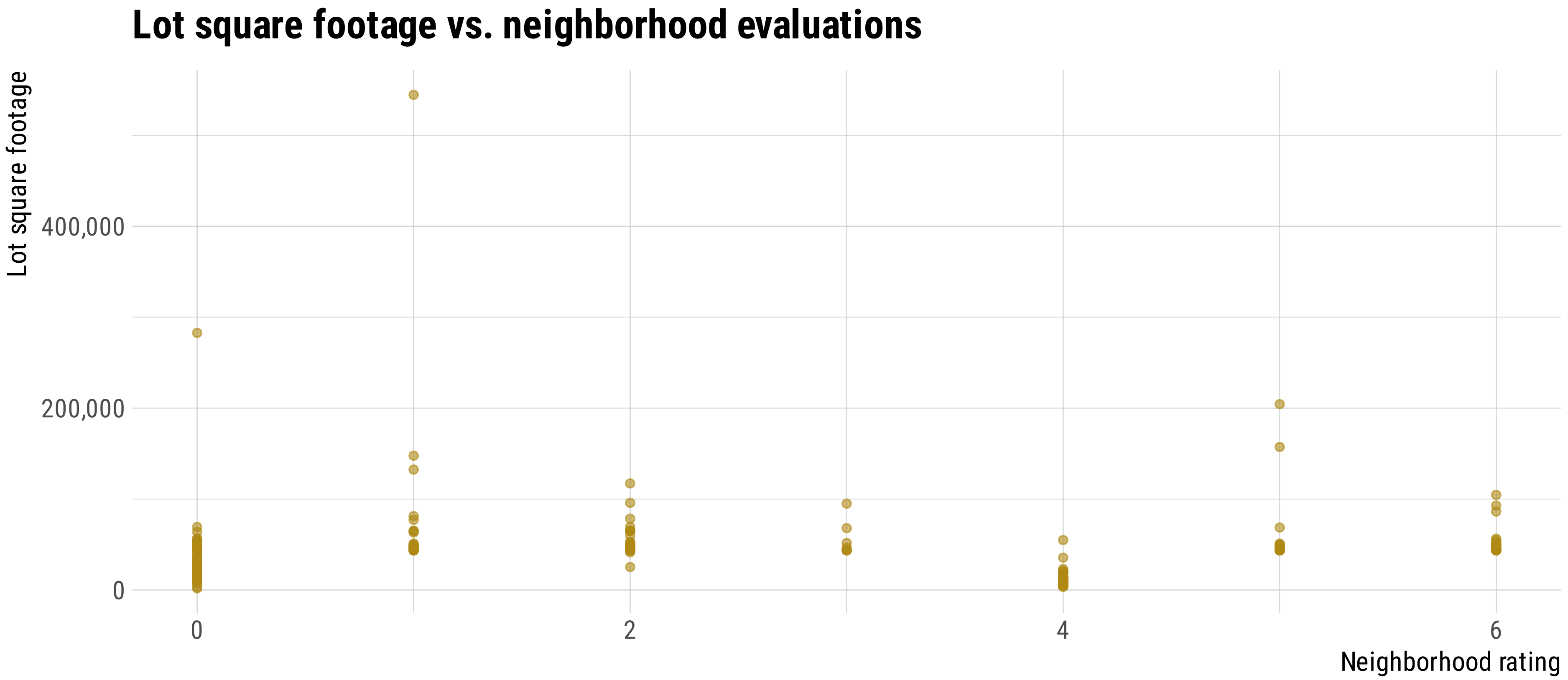
# Bivariate descriptive techniques



# Bivariate descriptive techniques



# Bivariate descriptive techniques



# Bivariate descriptive techniques

Let us start with the **covariance**.

The covariance gives two pieces of information about the *association* between two variables (say, $x$ and $y$): the **nature** and the **strength** of this relationship.

- **Population covariance** ($\sigma_{xy}$):

$$\sigma_{xy} = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N}$$

- **Sample covariance** ($s_{xy}$):

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

# Bivariate descriptive techniques

An **alternative** formula for the *sample covariance*:

$$s_{xy} = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}\right]$$

# Bivariate descriptive techniques

```r
data("smoke")                       # data from the "wooldridge" package.

smoke <- smoke %>% as_tibble()  # transforming it into a tibble.

smoke_filtered <- smoke %>%
  filter(cigs > 0)                  # what is this piece of code doing?

smoke_filtered %>%
  select(cigs, cigpric, educ, age) %>%
  head()
```

```
#> # A tibble: 6 × 4
#>    cigs cigpric  educ   age
#>   <int>   <dbl> <dbl> <int>
#> 1     3    57.7  12      58
#> 2    10    57.9  13.5    27
#> 3    20    60.3  12      24
#> 4    30    57.9  10      71
#> 5    20    60.1  12      29
#> 6    30    60.7  12      34
```

# Bivariate descriptive techniques

Data from `Mullahy (1997)`:

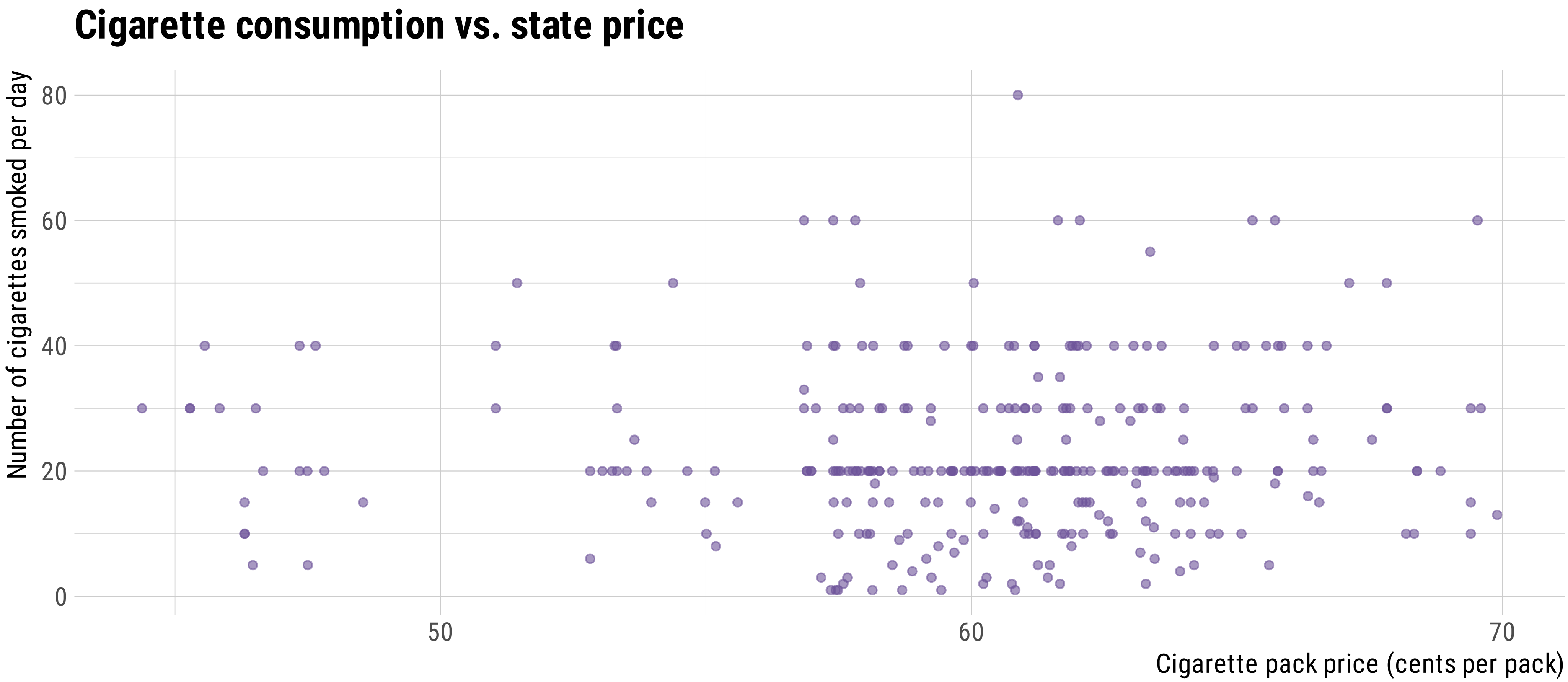
```
smoke_filtered %>%
  summarize(covariance_cigpric_cigs = cov(cigpric, cigs))
```

```
#> # A tibble: 1 × 1
#>   covariance_cigpric_cigs
#>                     <dbl>
#> 1                    1.75
```

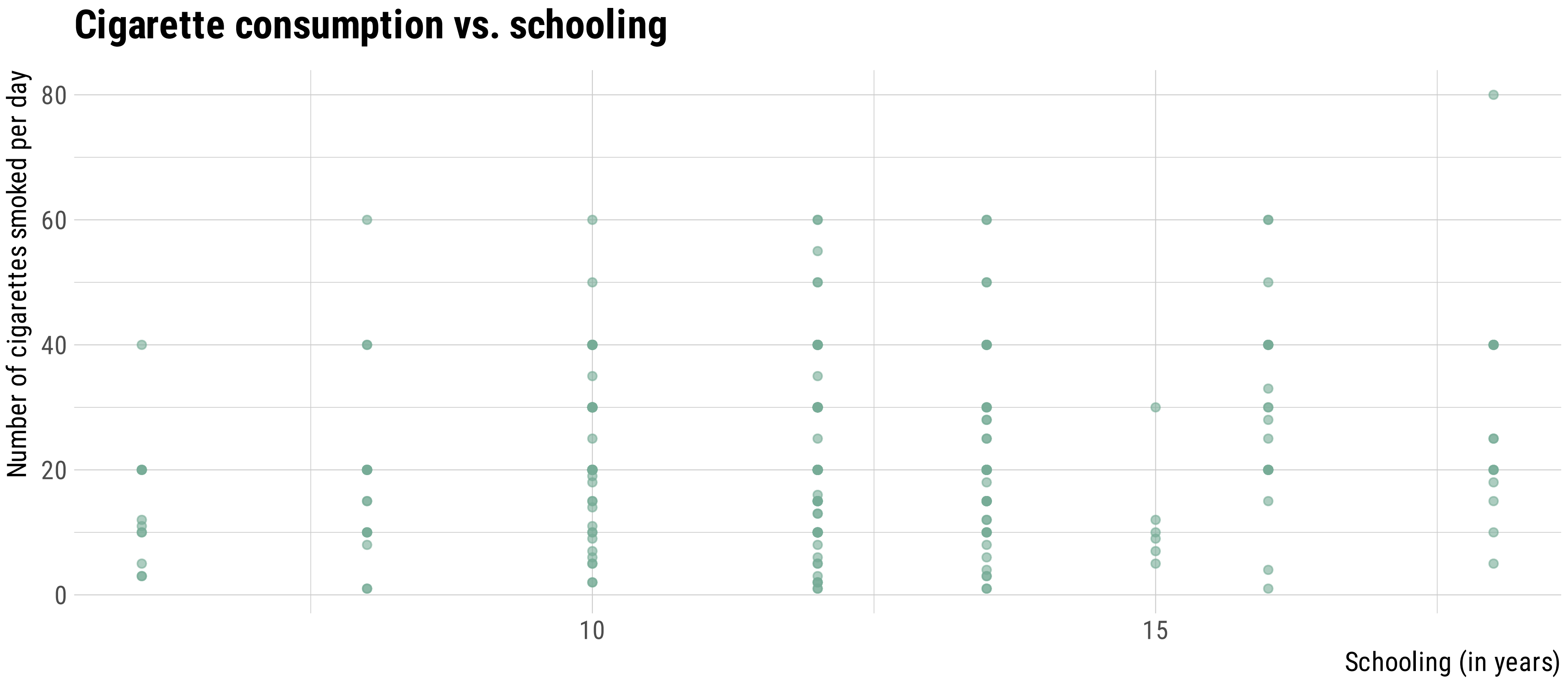
```
smoke_filtered %>%
  summarize(covariance_educ_cigs = cov(educ, cigs))
```

```
#> # A tibble: 1 × 1
#>   covariance_educ_cigs
#>                  <dbl>
#> 1                 5.43
```

# Bivariate descriptive techniques



# Bivariate descriptive techniques



# Bivariate descriptive techniques

Now, to the **correlation coefficient**.

The coefficient of correlation is *more specific* than the covariance.

The correlation coefficient implies a **linear relationship** between $x$ and $y$.

Therefore, in case the shape from a *scatter diagram* does not predict a **linear** relationship between the two variables, using the correlation may not be the best measure.

- **Population correlation** ($\rho$):

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

- **Sample correlation** ($r$):

$$r = \frac{s_{xy}}{s_x s_y}$$

# Bivariate descriptive techniques

The correlation formula relates the covariance between $x$ and $y$, divided by the interaction between their respective standard deviations.

One **advantage** of this coefficient relative to the covariance is that it lies between **-1** and **+1**.

- $r = -1 \Rightarrow$ *negative*, perfect linear relationship between $x$ and $y$;
- $r = +1 \Rightarrow$ *positive*, perfect linear relationship between $x$ and $y$;
- $r = 0 \Rightarrow$ *no* linear relationship between $x$ and $y$;

# Bivariate descriptive techniques

Data from `Mullahy (1997)`:

```
smoke_filtered %>%
  summarize(correlation_cigpric_cigs = cor(cigpric, cigs))
```

```
#> # A tibble: 1 × 1
#>   correlation_cigpric_cigs
#>                      <dbl>
#> 1                   0.0271
```

```
smoke_filtered %>%
  summarize(correlation_educ_cigs = cor(educ, cigs))
```

```
#> # A tibble: 1 × 1
#>   correlation_educ_cigs
#>                   <dbl>
#> 1                 0.156
```

# Bivariate descriptive techniques

Lastly, the **coefficient of determination**.

It is more widely known as the $R^2$ *coefficient*.

Given the *limitations* of the coefficient of correlation to precisely interpret values other than 0, -1, and +1, the coefficient of determination, $R^2$, can be **precisely** interpreted.

It is obtained by simply **squaring** the correlation coefficient (for either population or sample measures).

# Bivariate descriptive techniques

```
smoke_filtered %>%
  summarize(R2_cigpric_cigs = cor(cigpric, cigs)^2 * 100)
```

```
#> # A tibble: 1 × 1
#>   R2_cigpric_cigs
#>             <dbl>
#> 1          0.0733
```

```
smoke_filtered %>%
  summarize(R2_educ_cigs = cor(educ, cigs)^2 * 100)
```

```
#> # A tibble: 1 × 1
#>   R2_educ_cigs
#>          <dbl>
#> 1         2.45
```

# Data collection & sampling

# Data collection & sampling

At this day and age, **data availability** is part of our reality.

*But where do data come from?*

There are plenty of data collecting methods, and we will investigate *three* of them:

1. Direct observation;
2. Experimental methods;
3. Surveys.

# Data collection & sampling

**Direct observation**, as the name suggests, is the *simplest* method possible for collecting data.

The **experimental method** involves a random selection of subjects (individuals exposed to a treatment), with the sample being divided into two groups:

- The **control** group (*does not* take the treatment),
- The **treatment** group (*does* take the treatment).

Who has never been asked to participate in a **survey**?

# Data collection & sampling

Statistics is not free from *mistakes*, either voluntary or involuntary.

These can be summarized into two categories:

- *sampling* and
- *nonsampling* errors.

**Sampling** errors are discrepancies between sample statistics and population parameters, due to observations collected in the sample.

- Increasing the sample size ($n$) may help!

**Nonsampling** errors are more serious than the previous category, since increasing the sample size will hardly solve the problem.

- Selection *bias*!

# Next time: Descriptive Statistics in R, part II