Class 2d: Review of concepts in Probability and Statistics

Business Forecasting

Summarizing Data

Comparisions and Associations

Comparisions

- Descriptive and visual comparisons
- NOT declaring statistically significant differences, just eyeballing
- That's coming next

Comparing categorical variables

Do people living in rural areas are more likely to have diabetes?

- We have two categorical variables
- We can use frequency table to see how diabetes is distributed among the two types of areas:

	No Diabetes	Has Diabetes
Rural	8906	993
Urban	24780	3179

Comparing categorical variables

Do people living in rural areas are more likely to have diabetes?

- Are relative frequencies more helpful?
- Share of each subgroup within the sample

	No Diabetes	Has Diabetes	Total
Rural	0.24	0.03	0.27
Urban	0.65	0.08	0.73
Total	0.89	0.11	1.00

- Can we compare numbers in the *Has Diabetes* column?
- **Marginal frequencies** are total probabilities by group

Table of frequency

- We want to compare whether someone living in rural area is more likely to have diabetes than someone living in urban area
- So we want to see whether:

$$P(Diabetes_i = 1 | Area_i = Rural) > P(Diabetes_i = 1 | Area_i = Urban)$$

- We want to look at the **relative conditional frequencies**
- They are usually in **contingency tables**
  - Share with diabetes within urban sample
  - Share with diabetes within rural sample

	No Diabetes	Has Diabetes
Rural	0.90	0.10
Urban	0.89	0.11

$$P(Diabetes_i = 1 | Area_i = Rural) = \frac{P(Diabetes_i = 1 \cap Area_i = Rural)}{P(Area_i = Rural)} \approx \frac{0.03}{0.03 + 0.24} \approx 0.1$$

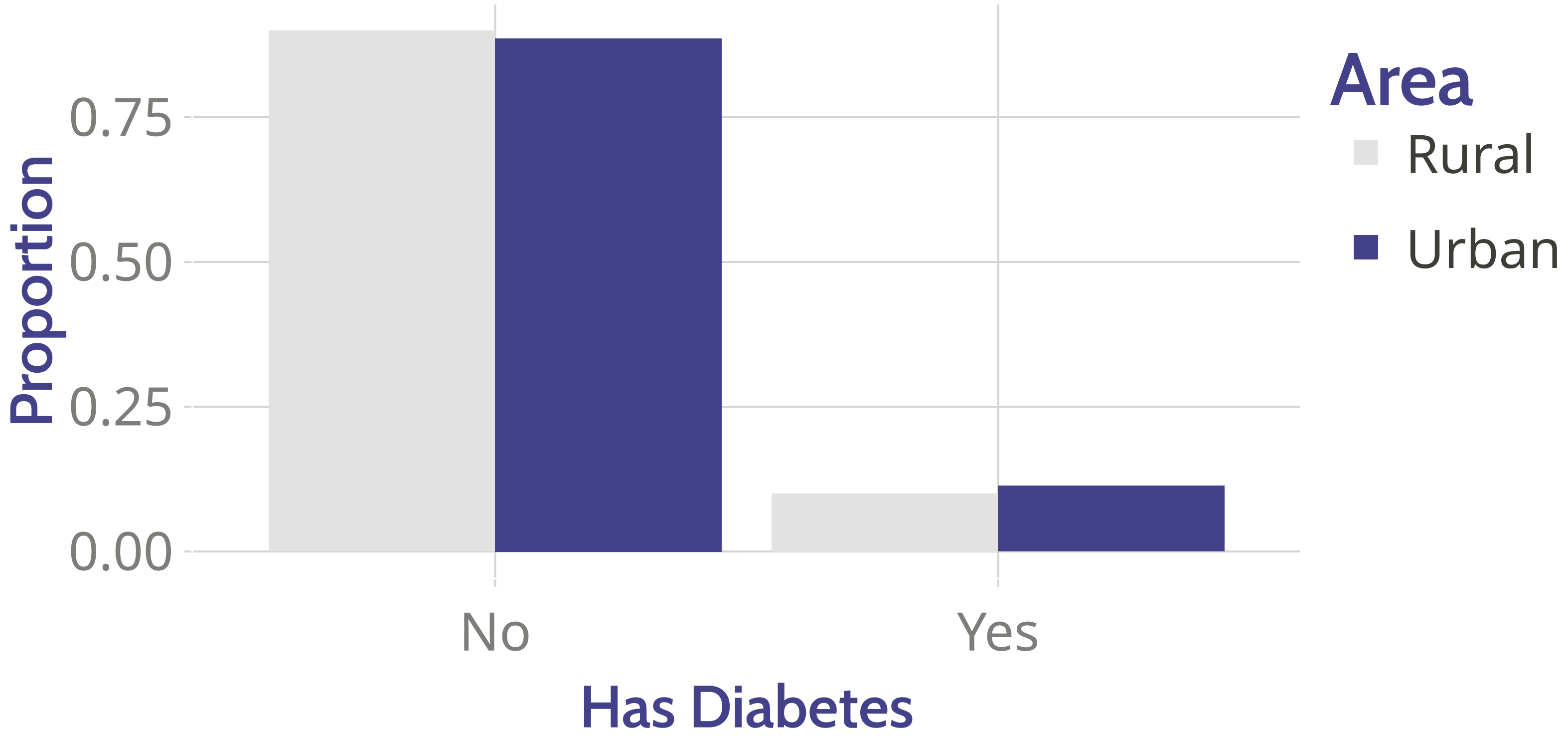
Or:

$$P(Diabetes_i = 1 | Area_i = Rural) = \frac{\text{Number live in Rural \& Have diabetes}}{\text{Number live in Rural}} = \frac{993}{993 + 8906} \approx 0.1$$

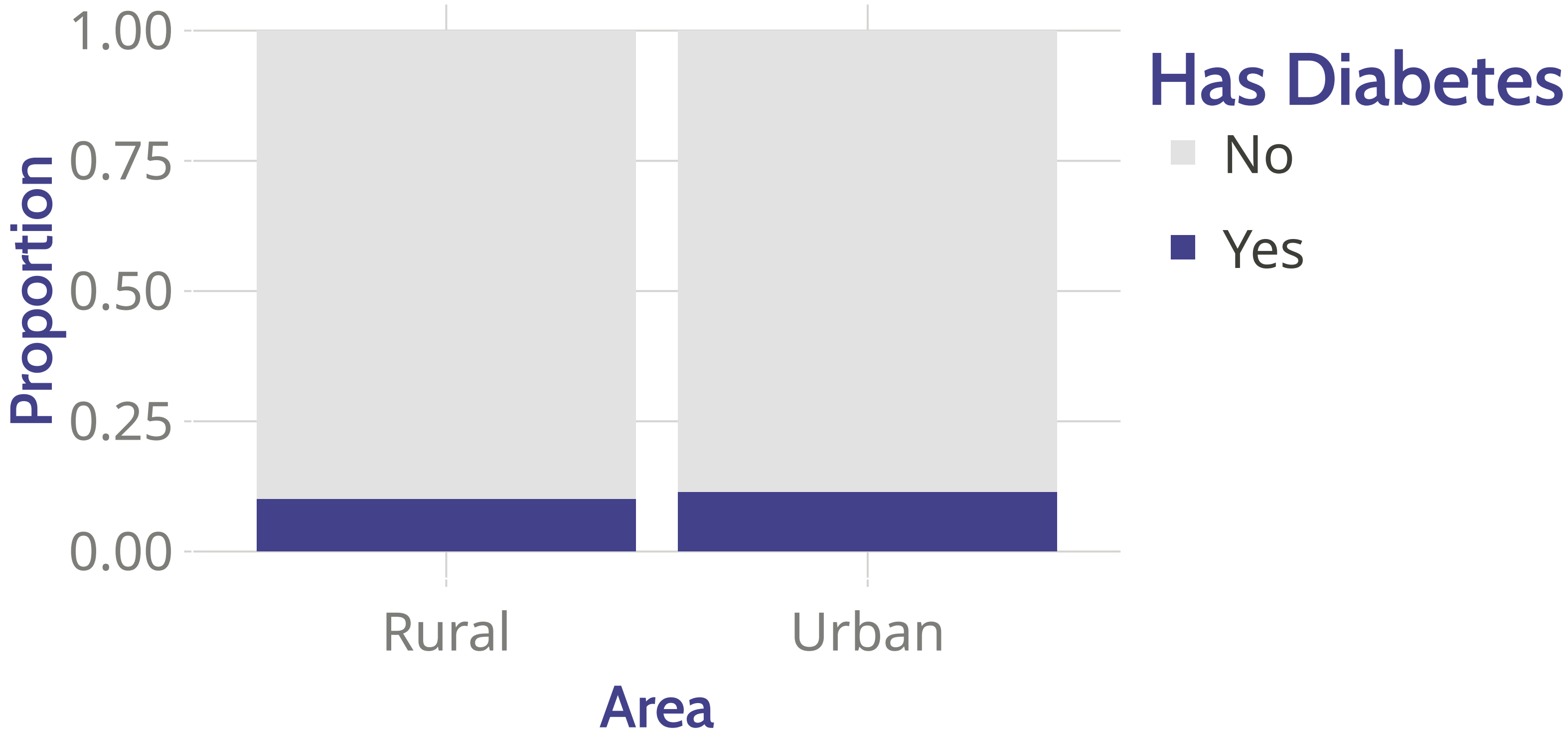
	No Diabetes	Has Diabetes
Rural	0.90	0.10
Urban	0.89	0.11

- What about marginal frequencies here?
  - Row sums should add up to 1
    - $P(Diabetes_i = 1|\text{Area=Rural}_i) + P(Diabetes_i = 0|\text{Area=Urban}_i)$
  - Column sums are meaningless
    - $P(Diabetes_i = 1|\text{Area=Rural}_i) + P(Diabetes_i = 1|\text{Area=Urban}_i)$

- We can visualize it on a barplot



- Or better on a **stacked barplot**



- *Stacked barplot* clearly shows the distribution of diabetes within each group

Practice

- Are you more likely to have diabetes if your mother had diabetes?
- By how much?

	No Diabetes	Has Diabetes
Mother No Diabetes	25270	2427
Mother Has Diabetes	8283	1721

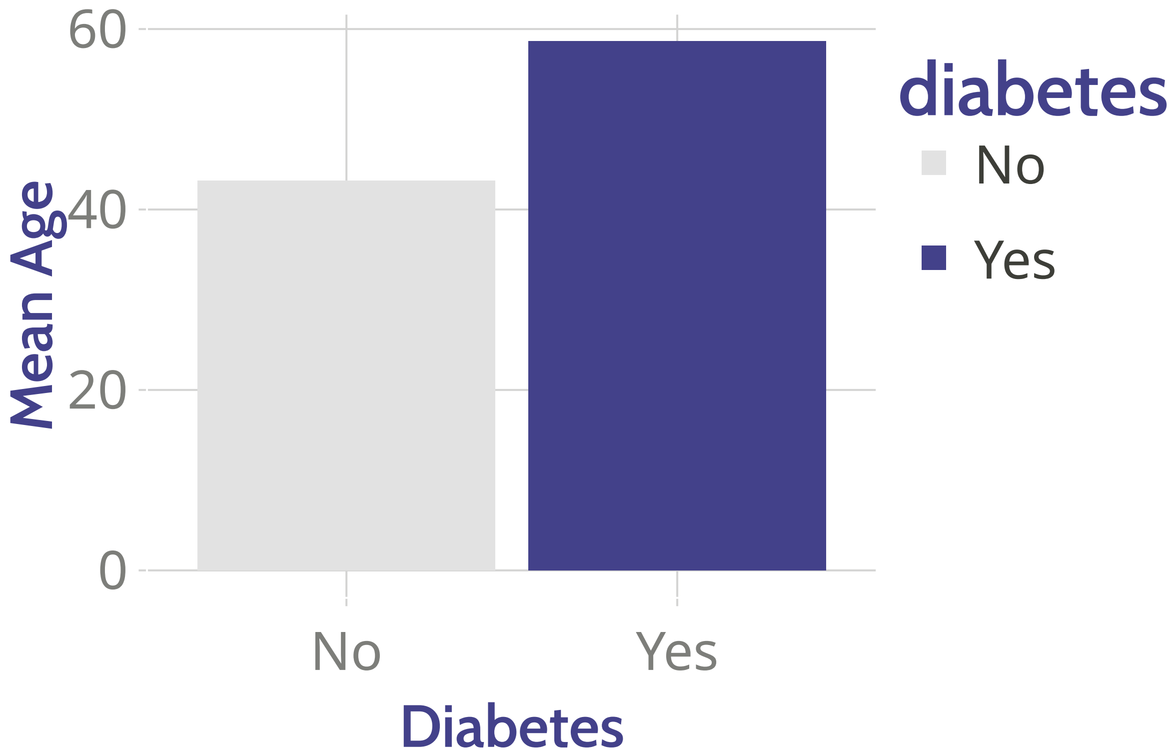
Practice

	No Diabetes	Has Diabetes
Mother No Diabetes	0.91	0.09
Mother Has Diabetes	0.83	0.17

- Does it mean that having diabetic mother **causes** higher change of having diabetes?

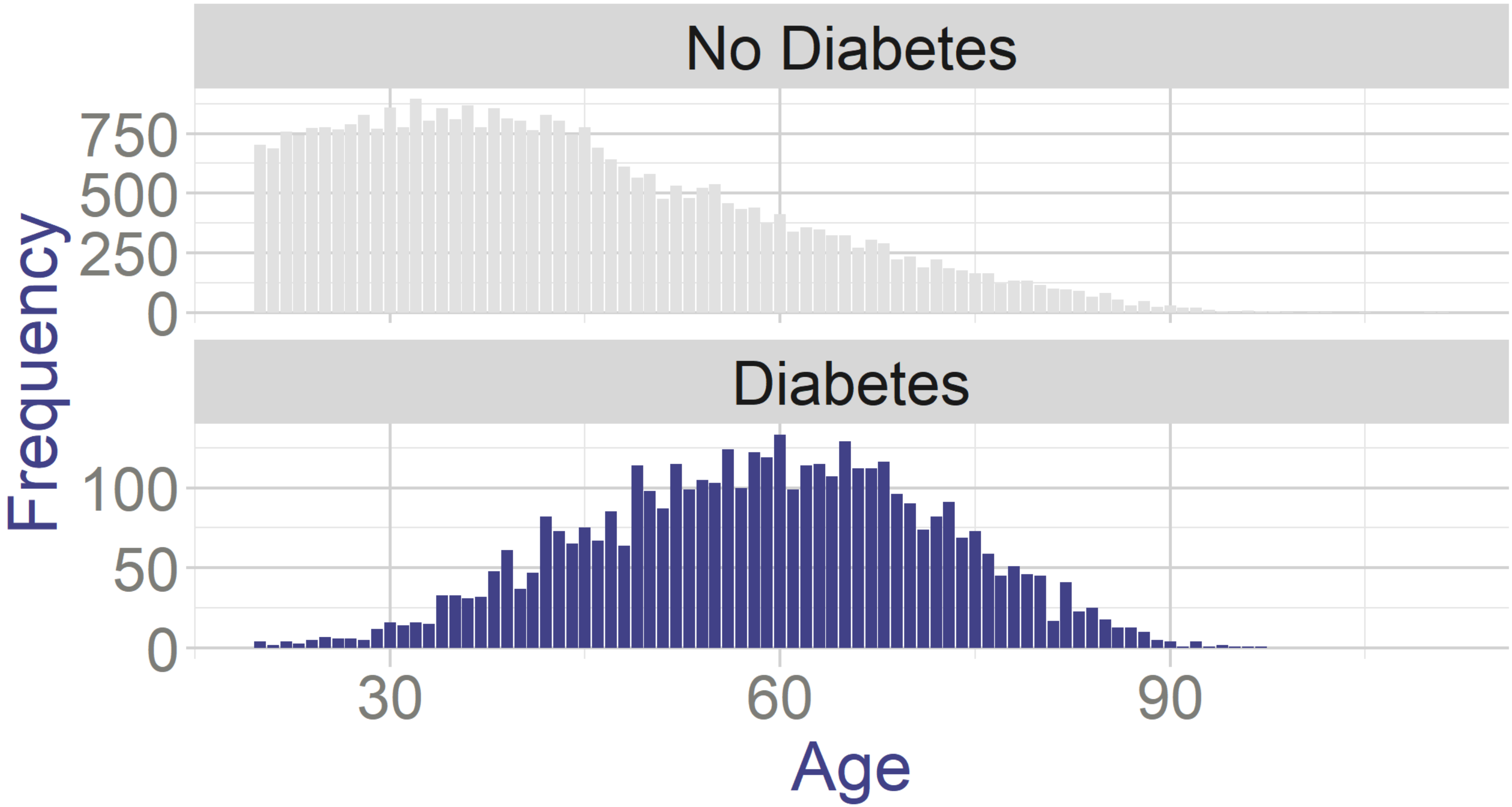
One quantitative and one categorical

- For quantitative variables we can compare some summary statistics
  - Are people with diabetes older than people without it?
  - *Example* means in two subpopulations



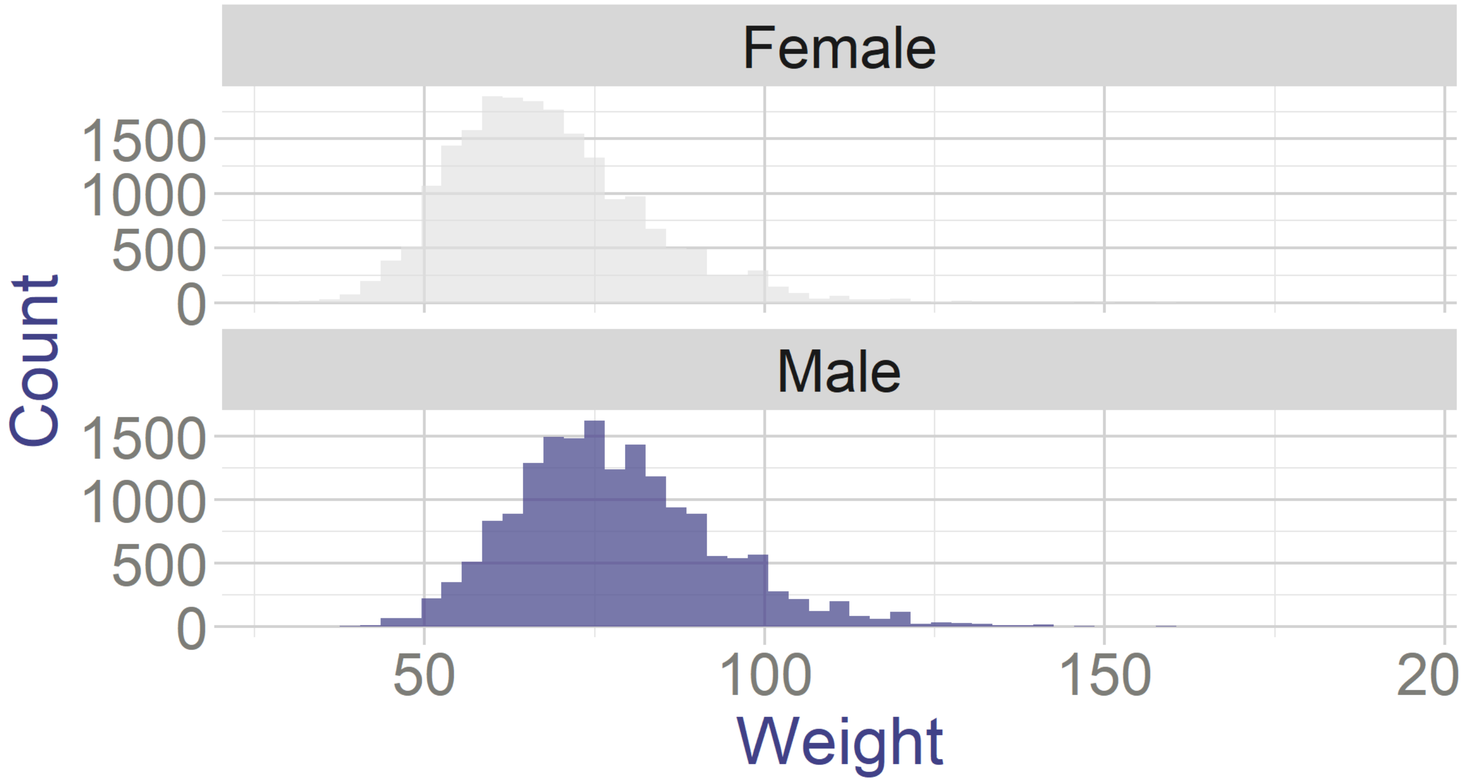
One quantitative and one categorical

- Or we can do Box and Whiskers plots as before
- Or we can compare the whole distributions of frequencies



One quantitative and one categorical

- For continuous variables we can use the same methods (except frequency distribution)
- Instead, we can compare densities or histograms
- Are men heavier than women?



Associations: Two Quantitative Variables

- Likely people would subscribe to the website to lose weight
- But do these people have resources?
- What is the relationship between Body Mass Index (BMI) and Income?
- More generally, how to measure association between two quantitative variables
- Association between qualitative variables is measured with contingency tables

Associations

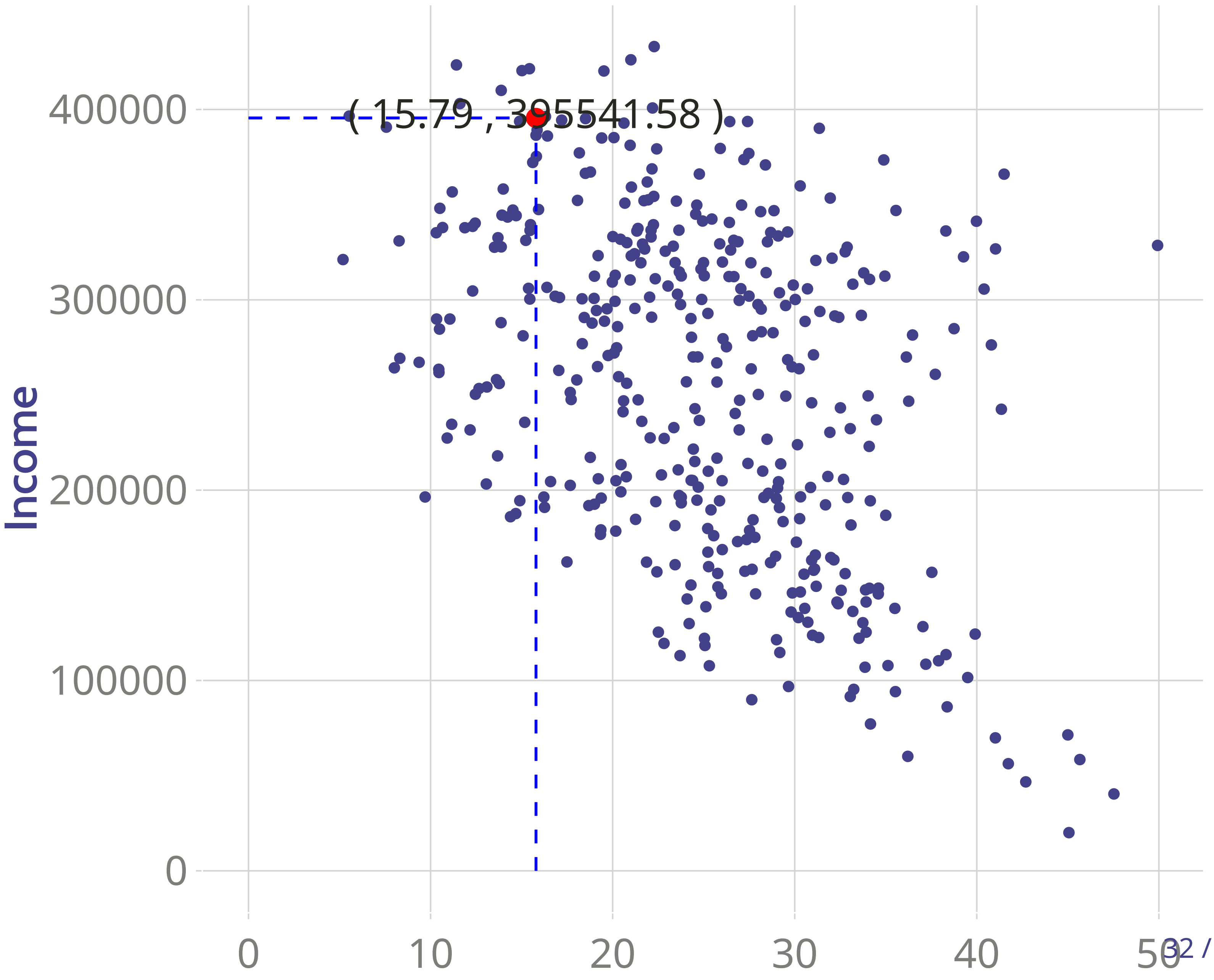
- Suppose we surveyed people from Guadalajara and CDMX about their BMI, education and income.
- Scatter plots show associations between two quantitative variables
  - We put variables of interest (*example*: Y and X) on the axis
  - We place observation on the cartesian plane using their values of variable X and Y: $\{(x_1, y_1), (x_2, y_2)..\}$
- In our case:
  - X axis is BMI
  - Y axis is Income
  - An individual i is placed on these axis based on $(BMI_i, Income_i)$

Show 4 entries

City	BMI	Education	Income
Mexico City	19.52	17.5	420224.44
Mexico City	22.16	15.3	368793.49
Mexico City	36.47	11.3	281512.52
Mexico City	24.56	13.4	344991.58

Showing 1 to 4 of 400 entries

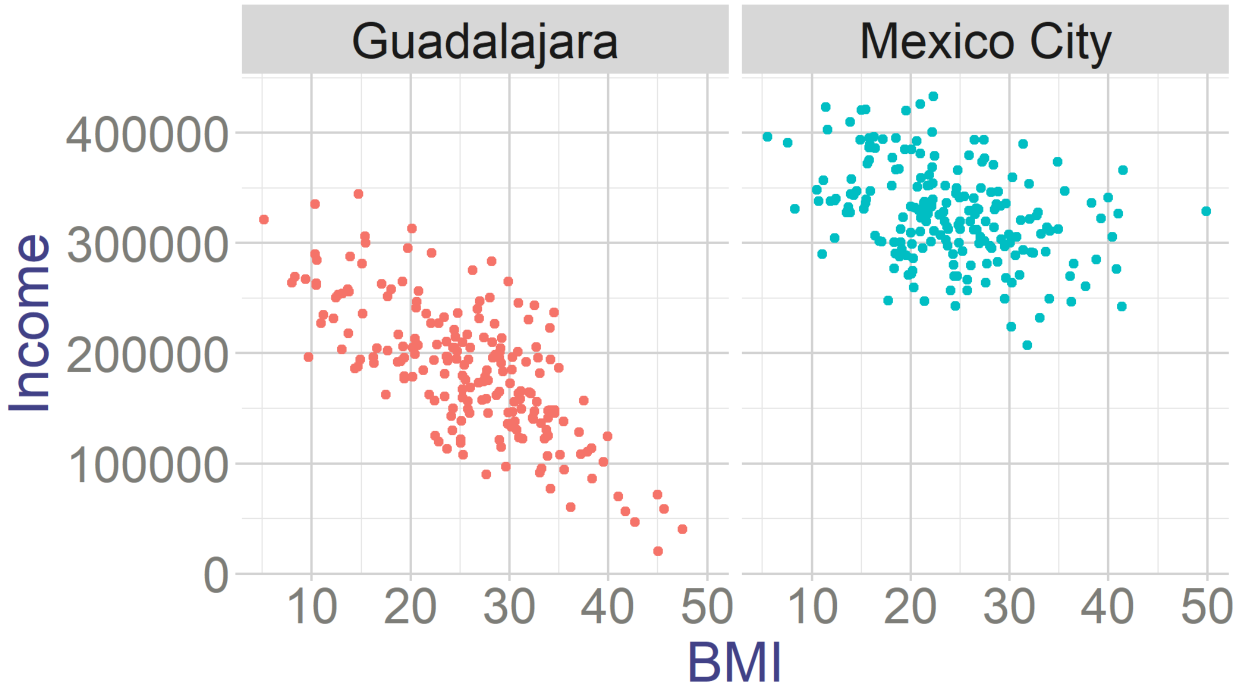
Previous 1 2 3 4 5 … 100 Next



(15.79 , 395541.58)
Income
400000
300000
200000
100000
0
0
10
20
30
40
50

Assocations

- Would you say that the relationship is stronger in Guadalajara or in Mexico City?



- How to measure the strength of the relationship?

Associations

Covariance

- **Covariance** measures the strength of the relationship between two variables.

$$\mathrm{Cov}(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)$$

And it's sample equivalent is:

$$\hat{\mathrm{Cov}}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

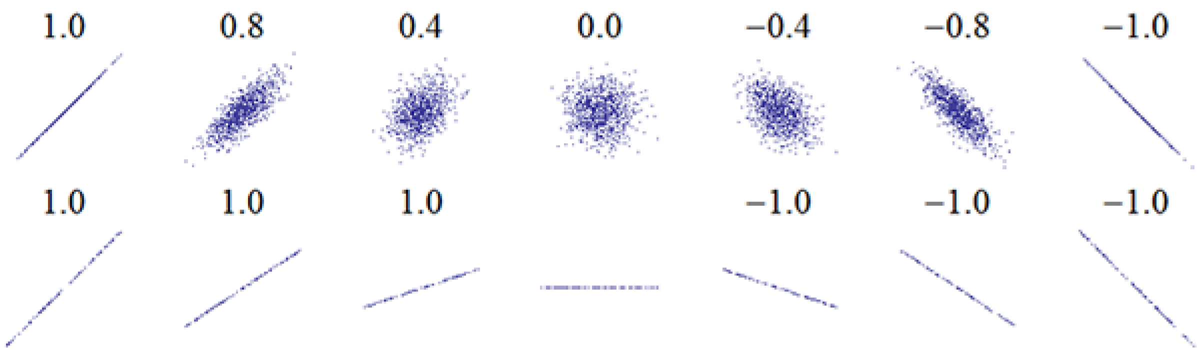
- Covariance whether the two variables move together

- Covariance increases when:
  - The relationship is stronger
  - The deviations of variables are larger

We use the Correlation coefficient to quantify the strength and direction of a relationship between two variables. $e.g.$, think about height and weight, or hours of sleep and irritability.

- The Pearson product-moment correlation coefficient is scale free and it ranges between -1 and 1.
- It is typically denoted by r, for sample data or by ρ (the greek symbol Rho), to indicate the population value.
- You have probably examined XY scatterplots to visualize this type of bivariate relationship, and have begun to evaluate the 2 dimensional attributes of the scattercloud to gain a sense of direction and strength of the relationship.

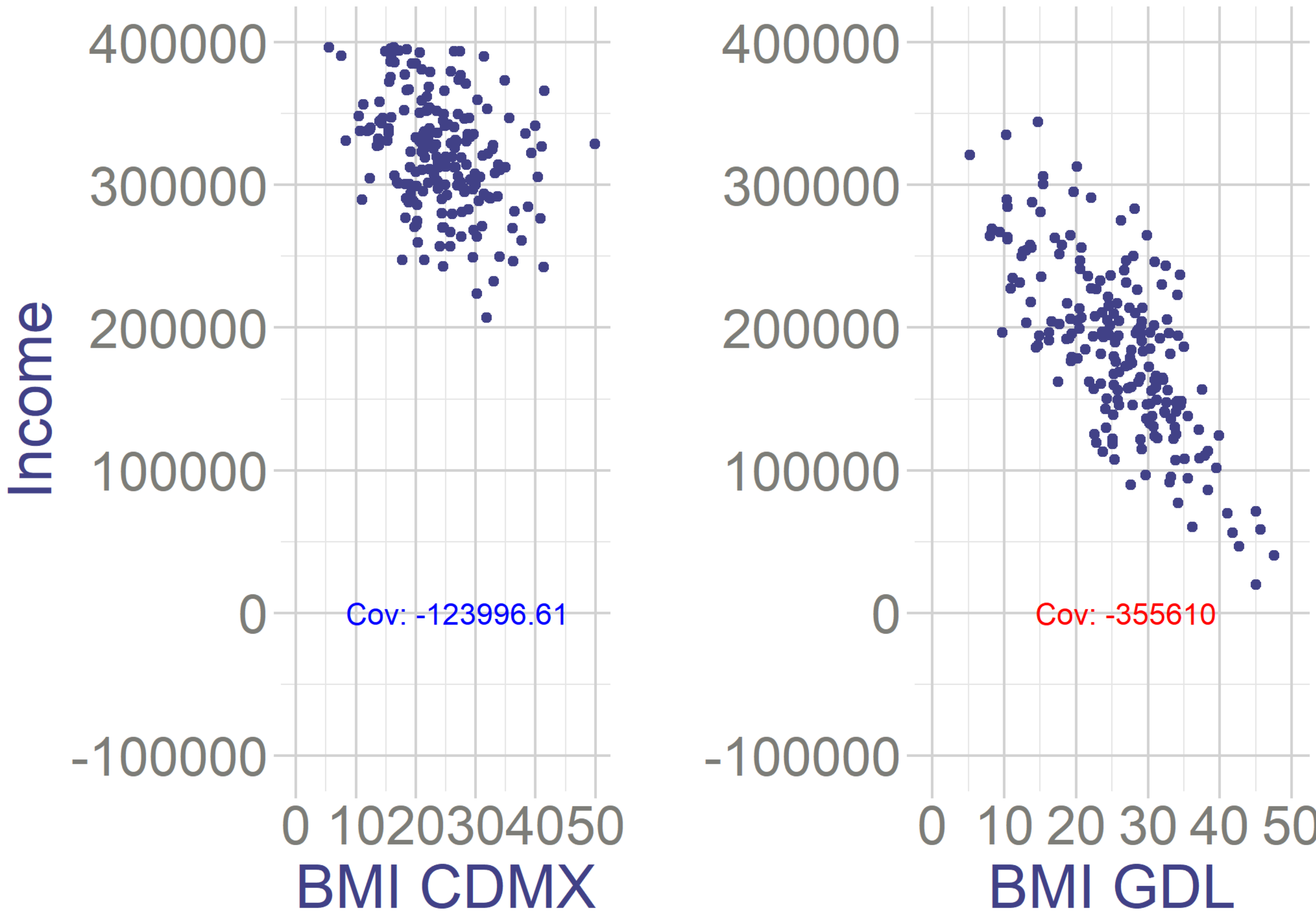
- Often, introductory textbooks show a figure like the following which depicts a series of XY scatterplots reflecting correlation patterns of differing size and sign. This one is the Wikipedia illustration.



- A correlation of -1 means that the X and Y variables have a perfect negative relationship and the data points fit a straight line with a negative slope.
- Similarly, a correlation of +1 means that X and Y have a perfect positive relationship and fall on a line with positive slope.

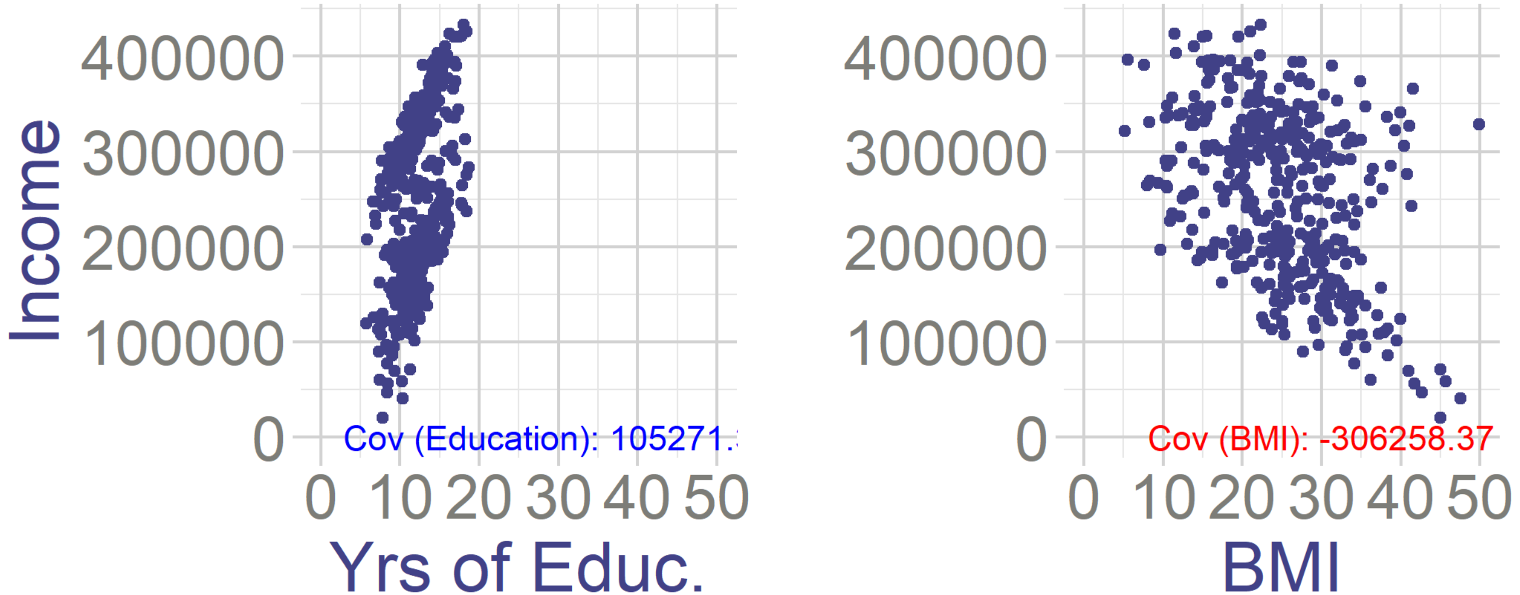
Source: https://shiny.rit.albany.edu/stat/rectangles/

Covariance


Income
400000
300000
200000
100000
0
-100000
Cov: -123996.61
0 10 20 30 40 50
BMI CDMX
400000
300000
200000
100000
0
-100000
Cov: -355610
0 10 20 30 40 50
BMI GDL

Covariance

- What has stronger relationship with Income: BMI or Years of Education?



- BMI has larger covariance
- But we can't compare covariances of different variables
- Covariance depends on the scales (or units) of the variable
- All else equal, larger standard deviation implies larger covariance
  - The squares are just bigger

Correlation

- **Correlation measures** the strength of a linear relationship between two variables.
- It ranges between -1 and 1

Population Correlation coefficient:

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$

Sample Correlation coefficient:

$$\hat{\rho}(X, Y) = \frac{\hat{\mathrm{Cov}}(X, Y)}{s_X \cdot s_Y}$$

Where $s_X = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$

Correlation

- Correlation is preferred over covariance because it's **scale-independent** and easier to interpret.
- Suppose that instead of measuring income (Y variable) in MXN , we measure it in Dollars.
  - Z income in dollars $Z = \frac{Y}{16}$
  - Is $Cov(X, Z) = Cov(X, Y)$?

$$
\begin{aligned}
cov(X, Z) &= \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu_X)(z_i - \mu_Z) \\
&= \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu_X)(\frac{y_i}{16} - \frac{\mu_Y}{16}) \\
&= \frac{1}{16} \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu_X)(y_i - \mu_Y) \\
&\neq cov(X, Y)
\end{aligned}
$$

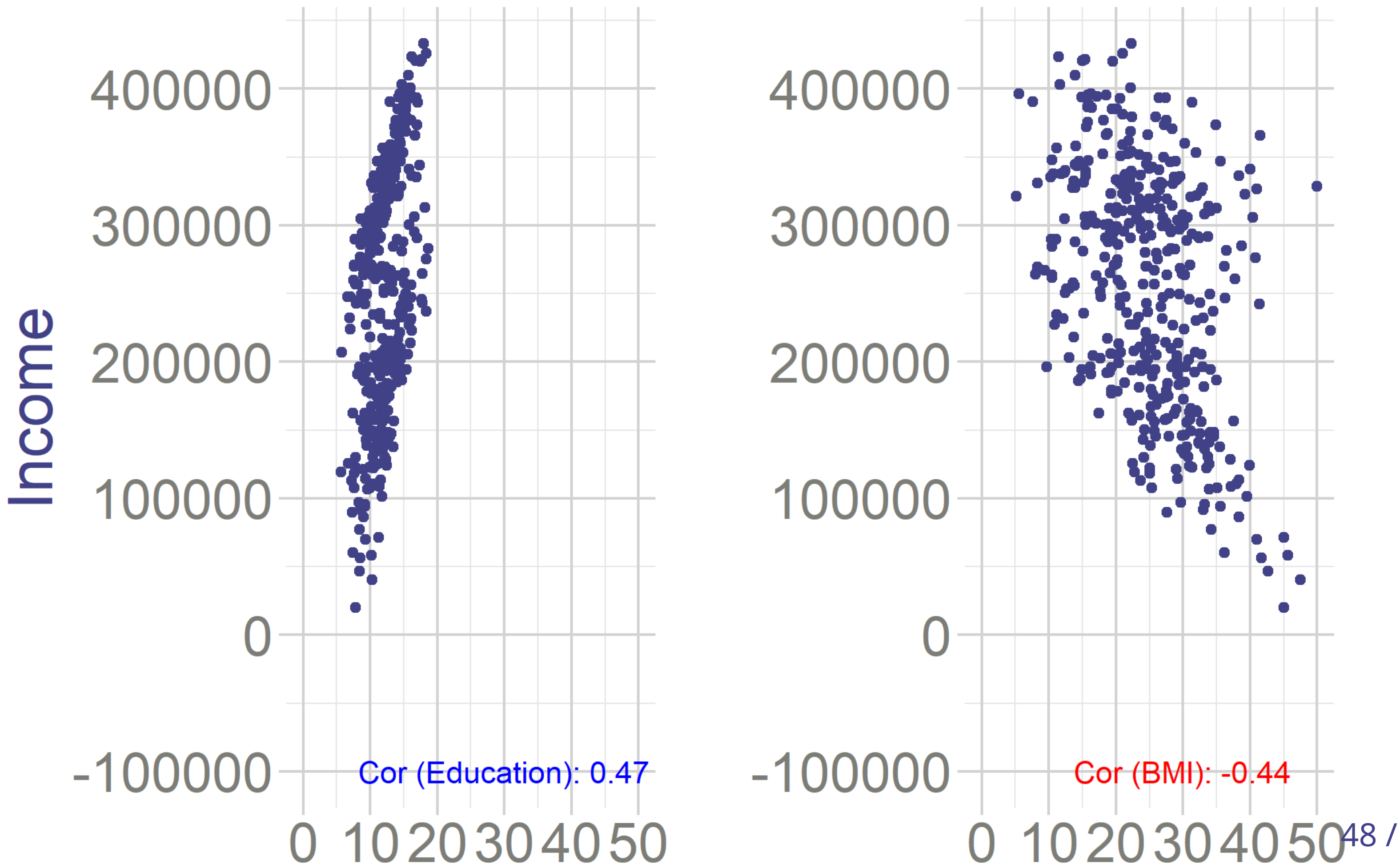
Correlation

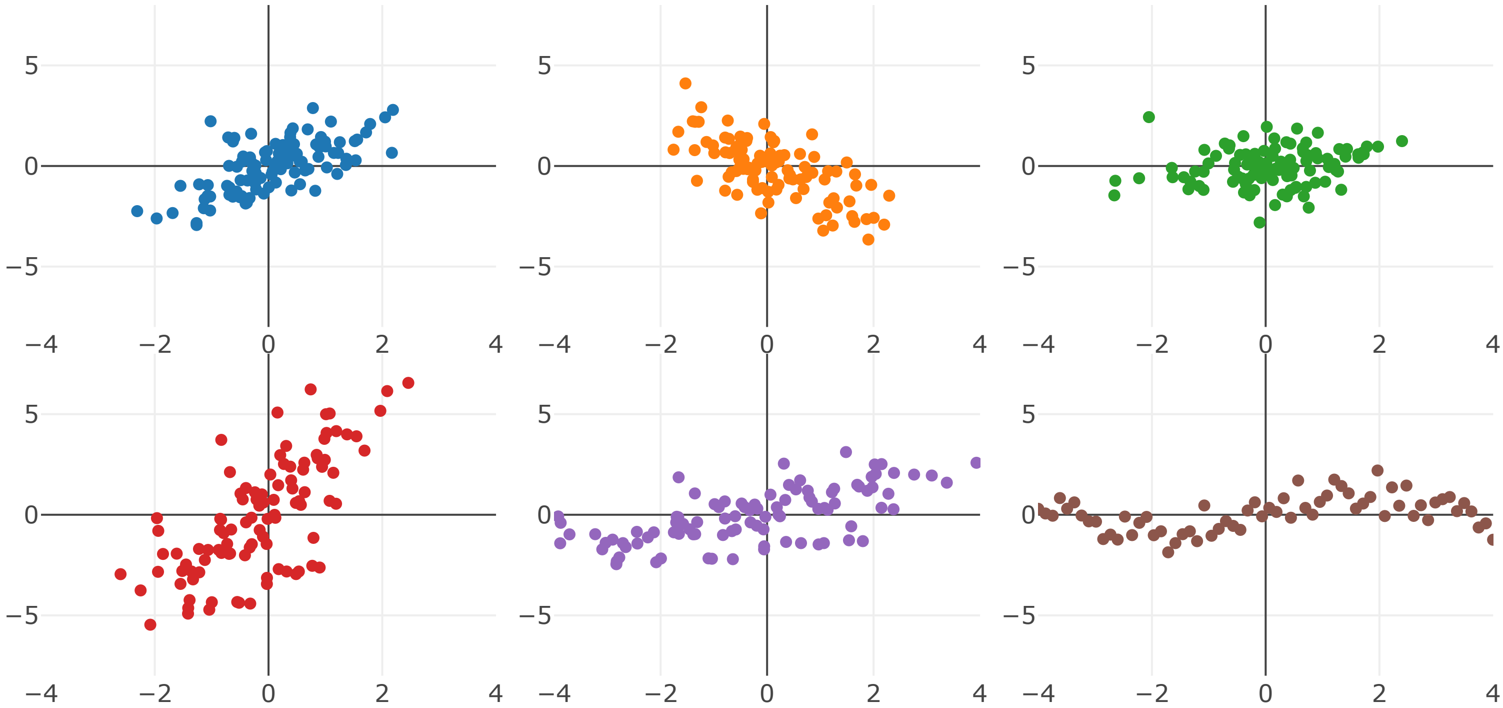
- Correlation is preferred over covariance because it's **scale-independent** and easier to interpret.
- Suppose that instead of measuring income (Y variable) in MXN , we measure it in Dollars.
  - Z income in dollars $Z = \frac{Y}{16}$
  - Is $\rho(X,Z) = \rho(X,Y)$?

$$
\begin{aligned}
\rho(X,Z) &= \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_X)(z_i-\mu_Z))}{\sqrt{\sum_{i=1}^{N}(x_i-\mu_X)^2}\cdot\sqrt{\sum_{i=1}^{N}(z_i-\mu_Z)^2}} \\
&= \frac{\frac{1}{N}\sum_{i=1}^{N}\sum_{i=1}^{N}(x_i-\mu_X)(\frac{y_i}{16}-\frac{\mu_Y}{16})}{\sqrt{\sum_{i=1}^{N}(x_i-\mu_X)^2}\cdot\sqrt{\sum_{i=1}^{N}(\frac{y_i}{16}-\frac{\mu_Y}{16})^2}} \\
&= \frac{\frac{1}{16}\frac{1}{N}\sum_{i=1}^{N}\sum_{i=1}^{N}(x_i-\mu_X)(y_i-\mu_Y)}{\frac{1}{16}\sqrt{\sum_{i=1}^{N}(x_i-\mu_X)^2}\cdot\sqrt{\sum_{i=1}^{N}(y_i-\mu_Y)^2}} \\
&= \rho(X,Y)
\end{aligned}
$$

Correlation

- Correlation with education is actually stronger



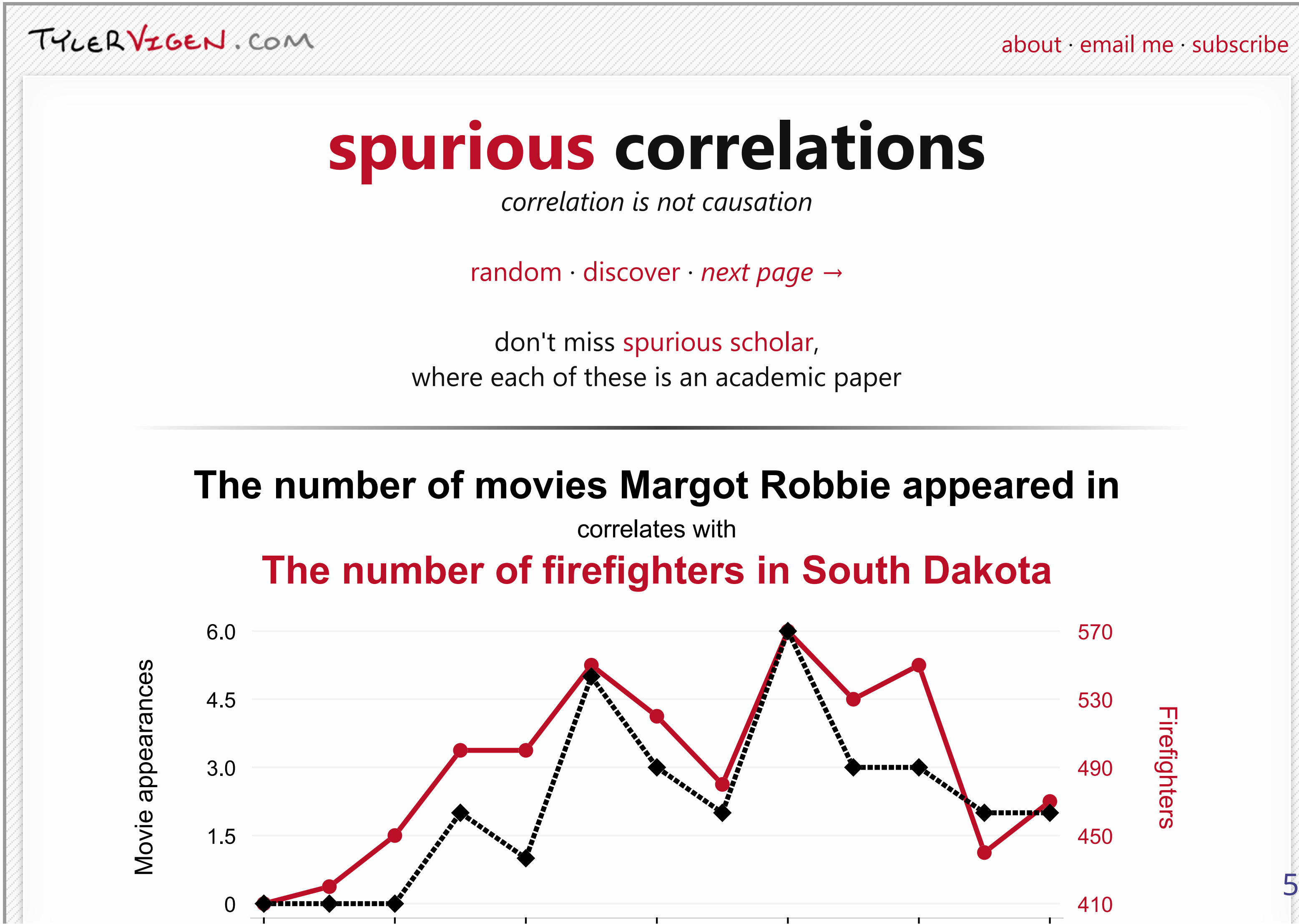

5
0
−5
−4
−2
0
2
4

Correlation

1. Correlation is a value between -1 and 1: $-1 \leq \rho(X, Y) \leq 1$.
2. Perfect positive correlation: $\rho = 1$. Perfect negative correlation: $\rho = -1$.
3. No linear correlation: $\rho = 0$, but this doesn't imply independence.
4. Correlation measures **linear** relationships; nonlinear relationships might not be accurately captured.
5. Correlation doesn't imply causation; a relationship could be coincidental.

Causality vs Correlation

Causality vs Correlation



Causality vs Correlation

- Less obvious examples
- You look at historical data from some media campaign
- You notice that people who were more exposed to ads were less likely to buy that product
- What can you conclude?
- Are people who were exposed to ads similar to people who were not?
- Maybe they were targeted in the first place because they are less likely to buy and you want to change it?

Causality vs Correlation

- Less obvious examples
- Education usually correlates with Income (correlation)
- Does it mean that if decide to get a degree, you will earn more? (causality)