universität wien

BAK3: Introduction to Quantitative Methods

Week 6: Bivariate Descriptive Statistics

Leonardo Carella

universität wien

The Plan for today

- Statistics: Measuring association between ***two*** variables:
  - Easy: **cross-tabulations**, **mean comparisons.**
  - Hard(er): **covariance** and the **correlation coefficient.**
- Coding in R:
  - All this stuff, plus **scatter plots**.

universität wien

Review Quiz

- What does this formula compute for the variable x of length n ?

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

- What would be the formula for the **variance**?
- Let's compute together the **standard deviation** of this variable:

$$x = \{-4,\ 0,\ 2,\ 2,\ 4,\ 8\}$$

universität wien

Cross-Tabulations

- Imagine you had a hypothesis about the association between two categorical (i.e. **nominal** or strictly **ordinal**) variables.
- For instance:
  - Presidential systems are more likely to experience coups.
  - Black applicants are less likely to be offered a job interview.
  - Men are more likely than women to participate in demonstrations.
- What are the **independent variables** in these hypotheses? What are the **dependent variables**?

universität wien

Cross-Tabulations

- Because the values in these variables have no mathematical meaning, there's not much fancy math we can do with them.
- Just to count the frequency of the possible outcomes on the dependent variable across each category of the independent variable.
- This will return a table (known as **cross-tabulation**, **cross-tab** or **contingency table**) with as many rows as there are possible values for the independent variable and as many columns as there are possible values for the dependent variable.

universität wien

Cross-Tabulations

Government System	Experienced Coup	Did Not Experience Coup
Parliamentary	8	42
Semi-Presidential	6	24
Presidential	18	22

- **Note**: I made up these numbers.

universität wien

Cross-Tabulations

Government System	Experienced Coup	Did Not Experience Coup
Parliamentary	16%	84%
Semi-Presidential	20%	80%
Presidential	45%	55%

Conditional Proportions: rows add up to 100%.

- **Note**: I made up these numbers.

universität wien

Means Comparison

- Imagine now you had a hypothesis about the association between a categorical independent variable and a numerical dependent variable.
- For instance:
  - Proportional electoral systems lead to a higher share of women in parliament relative to mixed-member or majoritarian systems.
  - Incumbents receive more votes than non-incumbent candidates.
  - Unemployed people support higher levels of welfare spending.

universität wien

Means Comparison

- We've already seen in the R sessions and in the assignments what we do in these cases.
- We compute the mean (or, less commonly, some other representative descriptive statistic) separately for observations in each group.

universität wien

Means Comparison

Electoral system	Number of countries (N)	Mean % women in parliament	Minimum	Maximum
Majoritarian	22	18.5%	5%	33%
Mixed	17	24.1%	8%	39%
Proportional	21	34.2%	17%	49%
Total	60	25.9%	5%	49%

- **Note**: I made up these numbers.

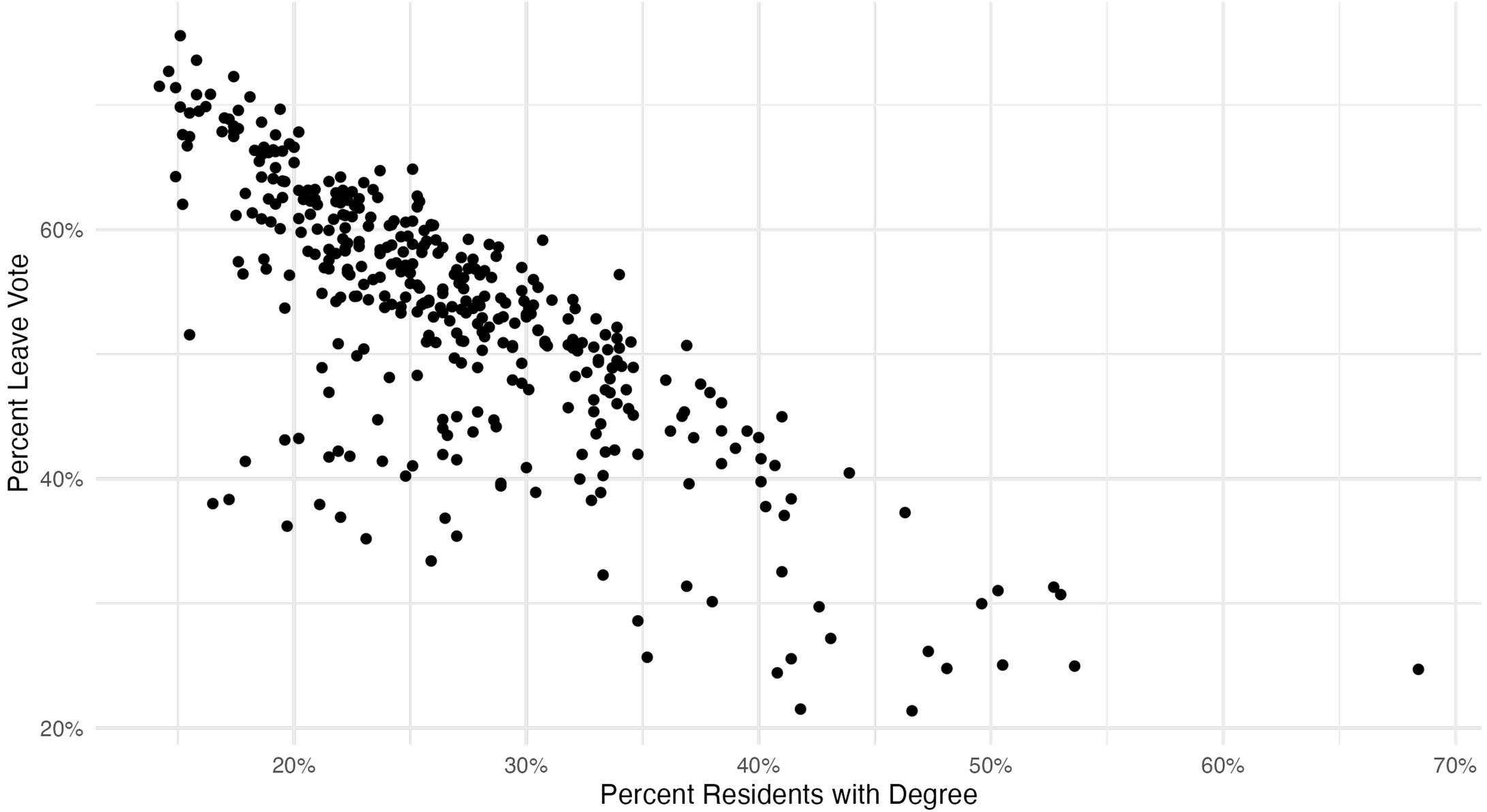
universität wien

Two Numerical Variables

- When you have a hypothesis linking two numerical variables, it's often useful to start with a visualisation: the **scatter plot.**
- The scatter plot shows observations as points in a space, with coordinates (x_i, y_i), where x_i is the value of the **independent** variable for observation i, and y_i is the value of the **dependent** variable.
- It's a convention (and not clear-cut in all cases), but remember to put the 'cause' variable on the x axis, and the 'effect' on the y axis.

universität wien

Some Scatter Plots



Covariance

- One way of quantifying direction (and strength) of associations is the **covariance**. The covariance of X and Y is given by…

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

- Can take **any value**. If positive, the association is positive. If negative, the association is negative.

universität wien

Covariance

- Why does this work? Imagine having only two people in your data: A weights 60kg and is 165cm; B weights 80kg and is 175cm.
- The mean weight is 70, the mean height is 170.

$$\mathrm{Cov}(W,H) = \frac{\left(w_A - 70\right)\left(H_A - 170\right) + \left(w_B - 70\right)\left(H_B - 170\right)}{n-1}$$

$$\mathrm{Cov}(W,H) = \frac{(-10)(-5) + (10)(5)}{2-1} = 50 + 50 = 100$$

- Weight and height are positively associated (groundbreaking!).

universität wien

Covariance

- Another basic example: person A works 10 hours a day, and has 2 hours of leisure time. Person B works 4 hours a day, and has 6 hours of leisure time.
- The mean work-time is 7, the mean leisure-time is 4.

$$\mathrm{Cov}(W, L) = \frac{\left(w_A - 7\right)\left(l_A - 4\right) + \left(w_B - 7\right)\left(l_B - 4\right)}{n - 1}$$

$$\mathrm{Cov}(W, L) = \frac{(3)(-2) + (-3)(2)}{2 - 1} = -6 + (-6) = -12$$

- Work-time and leisure-time are negatively associated.

universität wien

Covariance

- Order doesn't matter. $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$.
- The covariance is a close relative of the **variance**. In fact, $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$.
- ***Unlike the variance***, it can take positive **or negative** values. The sign refers to the direction of the association.
- Like the variance, it's **hard to interpret in terms of 'size'**. But it's the stepping stone to better measures of association (today: the **correlation coefficient**, week 11: the regression coefficient).

universität wien

Correlation Coefficient

- A more widely used measure of correlation is the Pearson correlation coefficient (a.k.a. Pearson's *r* or Pearson product-moment correlation coefficient, or just the correlation coefficient.

- It's the covariance of X and Y divided by the product of the standard deviation of X and the standard deviation of Y:

$$r_{X,Y} = \frac{\mathrm{Cov}(X, Y)}{s_X s_Y}$$

- It's a **unit-less** quantity. If you measure time in hours, minutes or seconds, if you measure weight in kgs, pounds or grams, it will still be the same.

universität wien

Correlation Coefficient

- Properties of Pearson's *r*:
  - It is always comprised between -1 and 1, where -1 is a perfect negative correlation, 1 is a perfect positive correlation an 0 is no correlation.
  - So we can interpret both **direction** (positive or negative) and **size** of the association (0.7 is larger than 0.2, whatever the units of your variables).
- BUT:
  - It only tells us about **linear** associations. Use scatter plots to detect if there are non-linear patterns in your data.
  - As usual, **correlation ≠ causation**, no matter how 'big' the coefficient.

universität wien

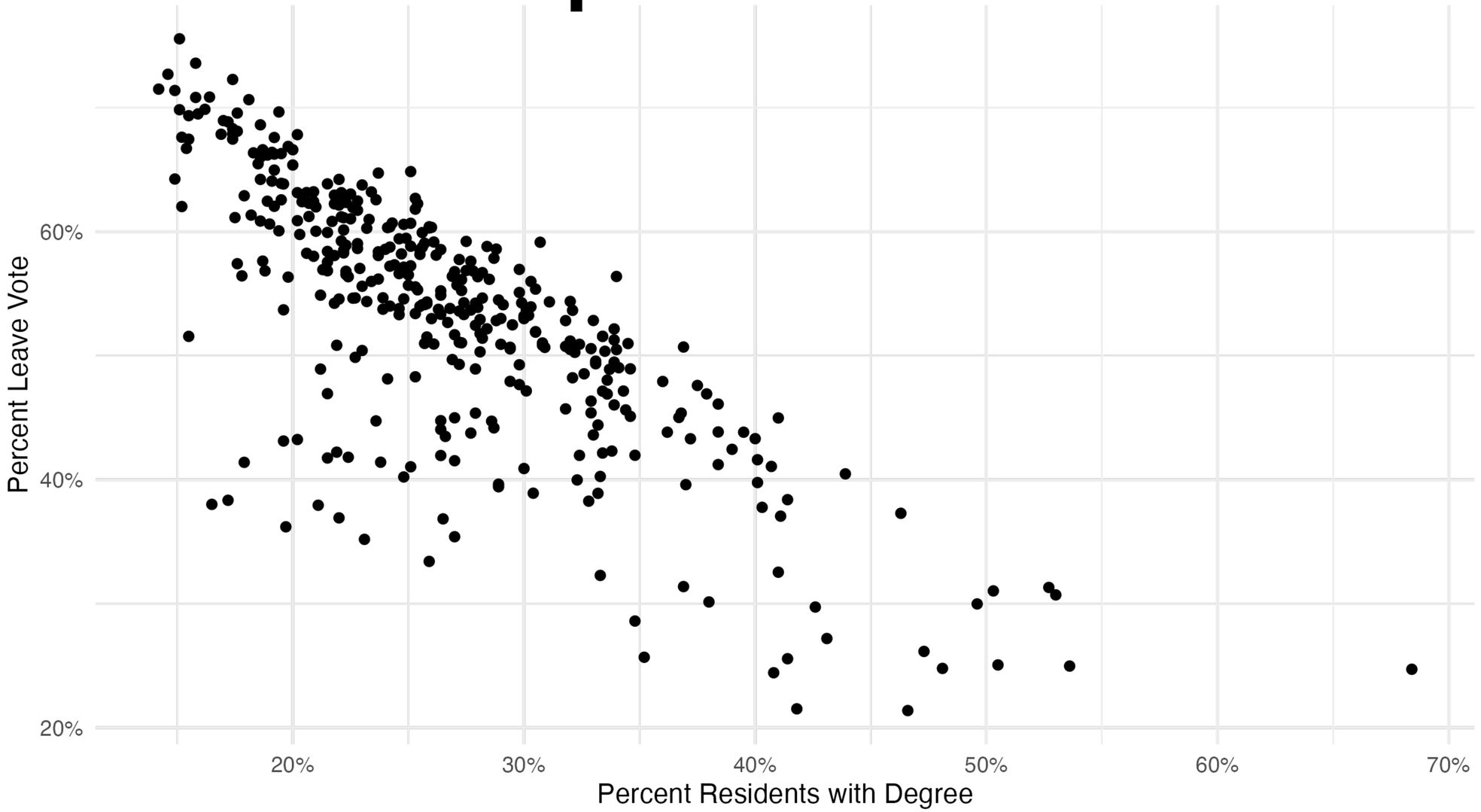
Examples

$r = -0.77$

$r = 0.91$

$r = 0.57$

$r = -0.36$



universität wien

Summing Up…

- When you have two categorical variables, show the association via a **cross-tabulation**.
- When you have one categorical variable and one numerical variable, show the association via a **means comparison**.
- When you have two numerical variables, measure the association with the **correlation coefficient** (show it with a **scatter plot**).
- And now, let's open RStudio…