

Chapter 12. Association

Statistical Association

- We are interested in the relationship between any two variables (Bivariate): weight and height, gender and attitude to domestic labor...
- Covariation: one variable goes with another variable in a particular direction or randomly...
- You have to think about measurement level again: different indicators for different level of measurement (look at the titles of chap. 13, 14, 15)
- First of all, make a bivariate table (crosstabulations) or make a scatter plot to get the big picture (the overall pattern).

Crosstabulation for categorical variables

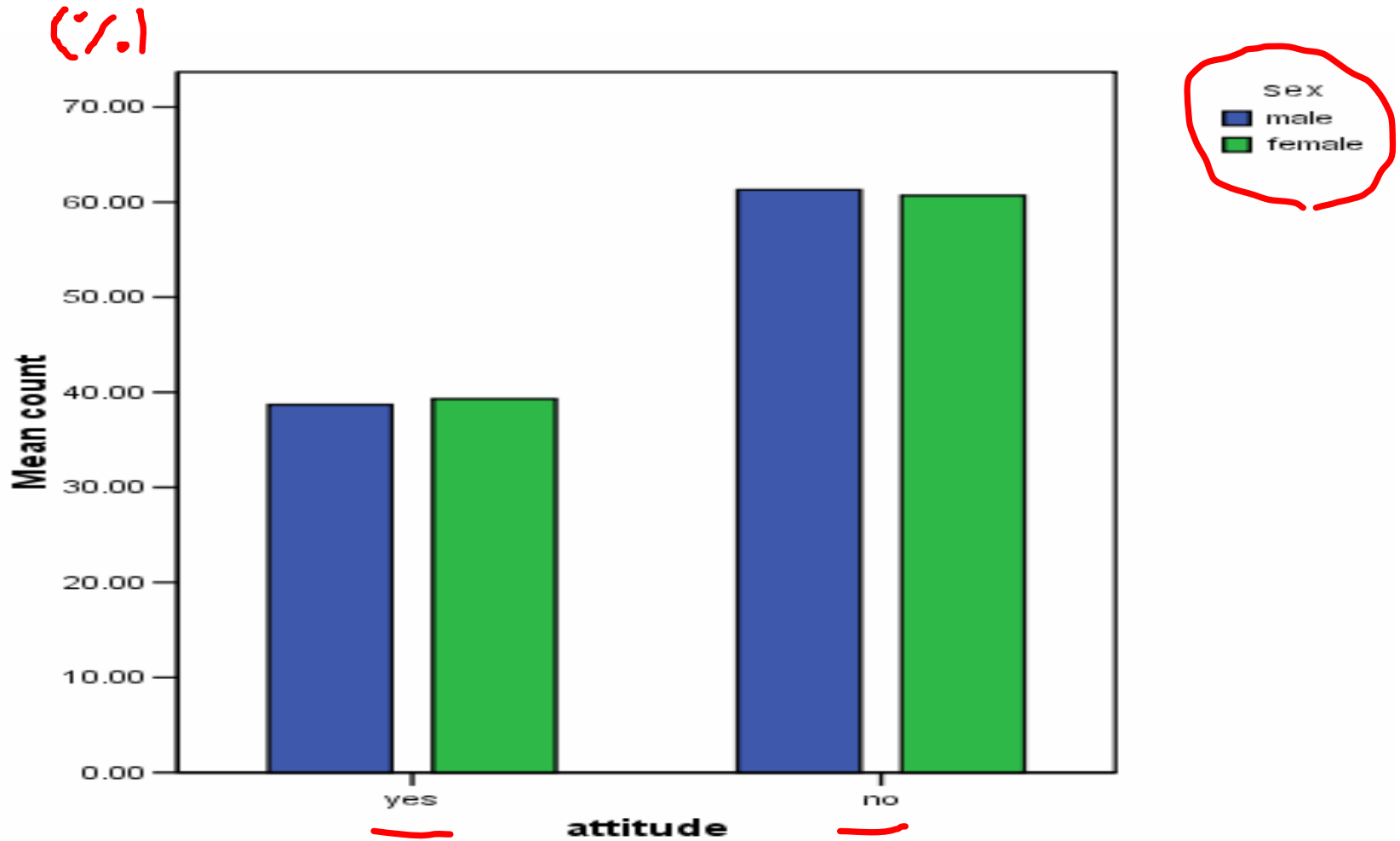
- You already know about this. (chi-square test)
- We are interested in the comparison of attitudes to abortion of men and women in GSS data.
- Data cleaning.
- Click “column percentages” because you want to compare the distribution of the dependent variable, the attitude, across sex.
- “The percentage is computed in which direction” is important to interpret the result.

Result

ABORTION IF WOMAN WANTS FOR ANY REASON ^ RESPONDENTS SEX Crosstabulation

			RESPONDENTS SEX		Total
			MALE	FEMALE	
ABORTION IF WOMAN WANTS FOR ANY REASON	Yes	Count	191	263	454
		% within RESPONDENTS SEX	38.7%	39.3%	39.1%
	No	Count	302	406	708
		% within RESPONDENTS SEX	61.3%	60.7%	60.9%
Total		Count	493	669	1162
		% within RESPONDENTS SEX	100.0%	100.0%	100.0%

No difference!



Interpretation


- Compare two distributions of “attitude” for male and female.
- Percentage calculated down the column! You should make comparisons in the opposite direction. That is, you should compare percentages across the row, “sex.”
- Whatever it is, yes or no, there does not seem to be any gender difference in abortion attitude.

Dependence and association

- Again, if there is a statistically significant relationship (association) between two variables, both are dependent.
- Otherwise, independent. You can't predict dependent variable's distribution with independent variable.
- What's your conclusion? Can you argue that there is statistical association between two variables here?
- It is also expected that you cannot reject the null hypothesis. Check it out.

Chi-square test

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.039 ^b	1	<u>.844</u>		
Continuity Correction ^a	.018	1	.892		
Likelihood Ratio	.039	1	.844		
Fisher's Exact Test				.855	.446
Linear-by-Linear Association	.039	1	.844		
N of Valid Cases	1162				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 192.62.

- This is why chi-square statistic is the most common measure of statistical association for categorical variables.

Another example

- Because your homework question 1 contains more than 2 categories for each variable, let me give you another example.
- Independent variable (column) is parents' social class; dependent variable (row) is offspring's level of education.
- Again, the direction in which percentages have been calculated is important. You have to make comparisons in the direction opposite to the direction in which percentages are computed.

	upper	middle	Lower
Less than HS	5	3	16
High school	15	12	23
College	10	36	41
Graduate	70	49	20
Total	100	100	100

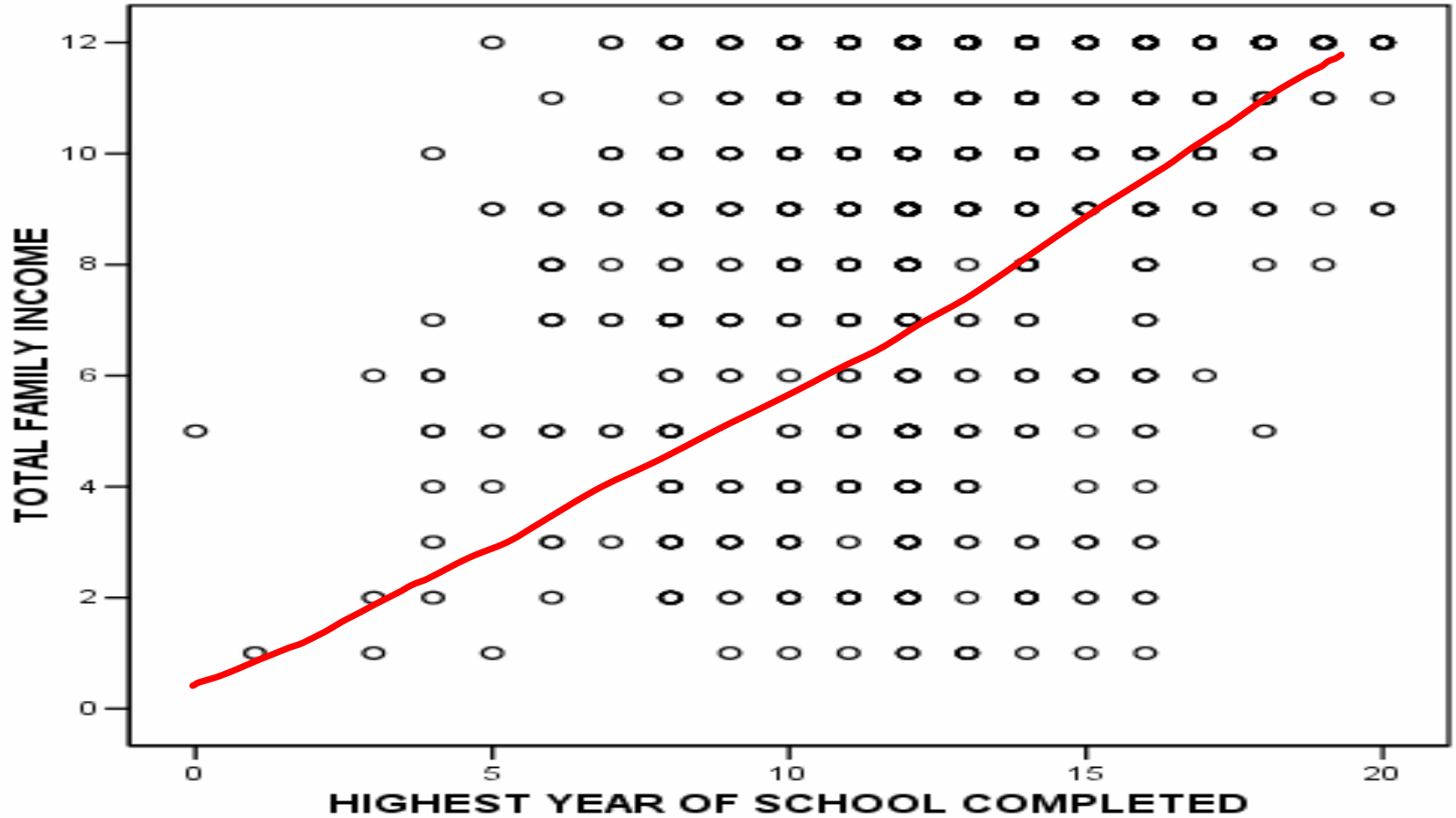
- Percentages given are calculated down the column. Therefore, you should think whether there is any big difference of each level of educational attainment among upper, middle, lower.
- You can say that two variables are strongly associated with each other since lower class is more likely to have lower level of education, while upper class is more likely to have higher level of education.

What if two variables are interval/ratio?

- You are interested in the relationship between “highest year of school completed” and “total family income” in GSS data.
- Data cleaning.
- The first step is a visual assessment by using scatter diagram.
- Go to Graphs> Scatter.

Result

(+)corr.



Interpretation

- It seems to me that there is significant (linear!) relationship between two variables, but the relationship is just moderately strong. (Pearson's correlation coefficient tells you about this exactly)
- Positive association (two variables covary in the same direction) or negative association (two variables covary in the opposite direction)?
- In this way, you have to think about three things. 1) statistical significance, 2) the strength of association, 3) direction of association.

Pearson's correlation coeff.

Correlations

		TOTAL FAMILY INCOME	RS HIGHEST DEGREE
TOTAL FAMILY INCOME	Pearson Correlation	1	<u>.319**</u>
	Sig. (2-tailed)	.	<u>.000</u>
	N	1364	1364
RS HIGHEST DEGREE	Pearson Correlation	.319**	1
	Sig. (2-tailed)	.000	.
	N	1364	1500

- FYI: The coefficient is 0.319. Significant at the level of 0.01(alpha)
- Even if there were non-linear relationship, the value of this coefficient could be almost 0.

** . Correlation is significant at the 0.01 level (2-tailed).