Quiz 5 (slides 5):

Assume that you have two categorical variables x and y and that variable x has J categories and variable y has K categories. Assume that each observation belongs to exactly one category with respect to x and that each observation belongs to exactly one category with respect to y. Assume that you have an iid sample, sample size n, from the corresponding bivariate distribution.

5 a) How do you construct the corresponding contingency table from the data?

5 b) Once you have the contingency table, how do you construct the corresponding relative frequency table and how do you calculate the corresponding marginal relative frequencies of the rows and the columns?

5 c) Consider the relative frequency that corresponds to the category j of the first variable x, and to the category k of the second variable y. This relative frequency is an estimate of certain probability. What probability?

5 d) How do you calculate the conditional relative frequency related to the category j of x conditioned on the category k of y and the conditional relative frequency related to the category k of y conditioned on the category j of x using the relative frequencies and the marginal frequencies?

5 e) The conditional relative frequencies from part d are estimates of certain probabilities. What probabilities?

5 f) How can you calculate estimated relative frequencies ("theoretical relative frequencies under independence") from the marginal frequencies if you assume independence?

5 g) How do you calculate the attraction-repulsion index related to the category j of x and the category k of y? If the index is larger than 1, what does it tell you? If the index is smaller than 1, what does it tell you?

Answers:

5 a) Contingency table is simply a table that displays observed frequencies. In this case, it has J rows and K columns and the element kj of the table is the number of observations that are in category j with respect to the first variable x and in category k with respect to the second variable y.

5 b) Relative frequency table is obtained from the contingency table by dividing each frequency in the contingency table by the sample size n. Marginal relative frequency of each row j is obtained by calculating the sum of the relative frequencies in the row j and marginal relative frequency of each column k is obtained by calculating the sum of the relative frequencies in the row k.

5 c) The probability P(x is in category j and y is in category k).

5 d) The conditional relative frequency related to the category j of x conditioned on the category k of y is equal to the relative frequency corresponding to the category j of x and the category k of y divided by the marginal frequency related to the column k. The conditional relative frequency related to the category k of y conditioned on the category j of x is equal to the relative frequency corresponding to the category j of x and the category k of y divided by the marginal frequency related to the row j.

5 e) The probabilities P(x is in category j assuming that y is in category k) and P(y is in category k assuming that x is in category j).

5 f) The estimated relative frequencies under independence ("theoretical relative frequencies under independence") related to the category j of x and the category k of y can be given as the product of the marginal relative frequency related to the row j and the marginal relative frequency related to the column k.

5 g) The attraction repulsion index related to the category j of x and the category k of y is equal to the corresponding relative frequency divided by the theoretical relative frequency under independence. If the value is larger than 1, we observe more than we should under independence. That is, there is positive dependence. If the value is smaller than 1, we observe less than we should under independence. That is, there is negative dependence.